

PERCEPTUALLY-BASED MODELLING FOR IMAGE COMPOSITE HARMONISATION

ALAN DOŁHASZ



A REPORT SUBMITTED AS PART OF THE REQUIREMENTS
FOR THE DEGREE OF PhD
AT THE SCHOOL OF COMPUTING AND DIGITAL TECHNOLOGY
BIRMINGHAM CITY UNIVERSITY, BIRMINGHAM, UK

AUGUST 2021

Supervisors:
Ian Williams, Carlo Harvey & Cham Athwal

Abstract

The field of image synthesis is concerned with generation of novel image content. Image compositing, the process of combining elements from existing image data into a seamless whole, is a common approach to image synthesis, employed in application domains such as visual effects in film, architectural visualisation, or augmented reality.

This thesis combines perceptual modelling with recent advances in machine learning in order to produce generalisable models of subjective visual realism in the context of digital image compositing. To achieve this, subjective visual realism in image composites is first modelled as a function of controllable local image transformations, applied to introduce composite-like distortions. These models are then validated and used to produce just-noticeable differences, describing average transformation magnitudes required for humans to distinguish such processed objects as unrealistic. The resulting models are then evaluated in the context of visual attention and refined in an image-wise fashion, before being approximated and generalised using deep learning techniques, particularly self-supervised transformation equivariant representation learning. The resulting models are subsequently shown to outperform baselines in an auxiliary task - image composite harmonisation, indicating that models trained on perceptual data are capable of generalising to related tasks.

Acknowledgements

Firstly, I would like to express my deep gratitude to Ian Williams and Cham Athwal for providing me with the opportunity to study for a PhD in the first place, as well as the many years of support, supervision, mentoring, encouragement and, most importantly, role modelling. I have learned a great deal from you, not only in the domain of research, and will always appreciate the time and effort you put in, despite my annoying character. Similarly, I would like to thank Carlo Harvey, whose insights, supervision, guidance and encouragement were instrumental in the timely completion of this work, as well as several related publications. I would also like to thank Zlatko Baracskaï - while it didn't work out in the end, you were the person to make me consider a PhD, for which I am grateful.

Additionally, I would like to thank my colleagues and friends at the Digital Media Technology Lab, who have inspired me or influenced my work in various ways over the years, namely: Sean Enderby, Ryan Stables, Sam Smith, Maite Frutos-Pascual, Nicholas Jillings, Dominic Ward, Jason Hockman, Matthew Cheshire, Carl Southall, Spydrion Stasis, Muadh Al-Kalbani, Chris Creed and Mattia Colombo.

Thanks to the developers and maintainers of open-source software, particularly Python, OpenCV, L^AT_EX, Tensorflow, PyTorch and countless other, who made much of this work possible.

Finally, I would like to thank my family, which I will do in Polish:

Chciałbym podziękować całej mojej rodzinie: Annie i Ryszardowi, jak również Edycie i Bartkowi, za wiarę we mnie, moje cele i marzenia, wsparcie i motywację w ich osiągnięciu, jak również za moje wychowanie i wartości które we mnie wszczepili. Prawie 15 lat na emigracji na prawdę pokazało mi jak duży wpływ na życie człowieka ma jego wychowanie i wartości. Cieszę się bardzo że właśnie z tego domu wyniosłem swoje.

Declaration

I confirm that the work contained in this PhD project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.

Signed
Alan Dolhasz

Date

Contents

Abstract	ii
Acknowledgements	iii
Declaration	iv
1 Introduction	1
1.1 Motivation	1
1.1.1 The Importance of Visual Information	1
1.1.2 The Fundamental Goals of Image Compositing	3
1.1.3 Perceptual Challenges of Image Compositing	5
1.1.4 Practical Challenges of Image Compositing	6
1.2 Application Domain	7
1.2.1 Visual Effects	7
1.2.2 Subjective & Perceptual Metrics	8
1.2.3 Image Forensics	8
1.2.4 Human Observer Modelling	8
1.3 Research Aim	9
1.4 Research Objectives	9
1.5 Methodology	9
1.6 Thesis Structure	10
1.7 Contributions	10
2 Understanding Perceptual Image Quality & Visual Realism	13
2.1 Introduction	13
2.1.1 Narrative Structure	14
2.2 Background on Human Visual Perception	15
2.2.1 Direct vs Inverse Problems	15
2.2.2 Computational Vision	17
2.2.3 Stages of Visual Processing	18
2.2.4 Scene Perception & Organisation	20
2.2.5 Colour Vision	21
2.2.6 Contrast Sensitivity	21

2.2.7	Visual Masking	22
2.2.8	Just-Noticeable Difference/Distortion	23
2.2.9	Visual Attention	23
2.2.10	Summary	25
2.3	Image Quality	25
2.3.1	Image Quality	25
2.3.2	Subjective Image Quality Assessment	28
2.3.3	Attributes Affecting Subjective Image Quality	28
2.3.4	Objective IQA	29
2.4	From Image Quality to Visual Realism	30
2.4.1	Limitations of Representation	31
2.4.2	Definitions of Visual Realism	31
2.4.3	Perception of Visual Realism	33
2.4.4	Image Features Affecting Realism	35
2.5	Methods for Subjective IQA	36
2.5.1	Subjective IQA Measurement	36
2.5.2	Stimulus Presentation	36
2.5.3	Grading Scales	38
2.5.4	Experimental Conditions & Presentation	40
2.5.5	Analysis of Results	43
2.5.6	Other Comparative Frameworks	43
2.6	Methods for Objective IQA	44
2.6.1	Signal Fidelity Measures	44
2.6.2	Perceptual Visual Quality Metrics (PVQM)	45
2.7	Methods for Measuring Realism	46
2.7.1	Realism & Visual Attention	51
2.8	Summary	52
3	Advances in Machine Learning for Image Quality	54
3.1	Introduction	54
3.2	Background on Machine Learning	54
3.2.1	Introduction	54
3.2.2	The General Learning Problem	55
3.2.3	Maximum Likelihood Estimation	57
3.2.4	Types of learning algorithms	58
3.2.5	Gradient-based Optimisation	59
3.2.6	Stochastic Gradient Descent	60
3.2.7	Generalisation	61
3.2.8	No Free Lunch Theorem & Regularisation	62
3.2.9	Hyperparameters	63
3.2.10	The essence of a machine learning algorithm	63
3.3	Deep Feedforward Networks	64

3.3.1	The Neuron	64
3.3.2	Activation Function	65
3.3.3	Cost Functions	67
3.3.4	The Universal Approximation Theorem	68
3.3.5	Convolutional Neural Networks	68
3.3.6	Properties of Convolutional Layers	70
3.3.7	Other Neural Network Layers	70
3.4	Architectures & Applications of CNNs	71
3.4.1	Image Classification	71
3.4.2	Object Detection	72
3.4.3	Semantic Segmentation	72
3.4.4	Image Synthesis	73
3.4.5	Perceptually-based Tasks	73
3.5	Limitations of Existing Approaches	79
3.6	Summary	80
4	Modelling Perceptual Realism in Image Composites	81
4.1	Introduction	81
4.2	Background & Related Work	82
4.2.1	Importance of Visual Realism	82
4.2.2	Image Features Affecting Realism	82
4.2.3	Synthetic Image Composites	83
4.2.4	Difference Thresholds	84
4.2.5	Towards Generalisable Models	84
4.3	Methodology	85
4.3.1	Overview & Motivation	85
4.3.2	Experimental Design	86
4.3.3	Synthetic Composite Generation	86
4.3.4	Composite Feature Selection	86
4.3.5	Base Image Dataset	92
4.3.6	Dataset Statistics	92
4.3.7	Apparatus & Task	94
4.3.8	Observers	100
4.3.9	Analysis of Results	100
4.4	Results	102
4.4.1	Goodness-of-fit Evaluation	102
4.4.2	Psychometric Functions & JNDs	102
4.4.3	Qualitative Evaluation	105
4.5	Discussion	110
4.5.1	Overview	110
4.5.2	Observer Performance & Lapse Rates	110
4.5.3	Qualitative Analysis	111

4.5.4	Summary	119
4.6	Conclusions	120
5	Impact of Attention, Task & Feature Type	121
5.1	Introduction	121
5.2	Related Work	122
5.2.1	Visual Attention	122
5.2.2	Modelling VA	125
5.2.3	Applications of VA Models	127
5.2.4	Gaze Tracking for VA Modelling	127
5.2.5	Visual Tasks & Gaze Metrics	128
5.2.6	Application of Gaze Metrics	130
5.2.7	Summary	131
5.3	Methodology	131
5.3.1	Overview	131
5.3.2	Stimuli	133
5.3.3	Observers	135
5.3.4	Apparatus	135
5.3.5	Procedure	135
5.3.6	Analysis	138
5.3.7	Qualitative Analysis of Fixation Maps	141
5.3.8	Hypotheses	141
5.4	Results	142
5.4.1	Realism Ratings	143
5.4.2	Fixation Counts	143
5.4.3	Fixation Durations	146
5.4.4	Time to First Fixation on Object	146
5.4.5	Duration of First Fixation on Object	147
5.4.6	Area of Interest Similarity	147
5.4.7	Inter-Observer Consistency	149
5.4.8	Response Time	149
5.5	Discussion	152
5.5.1	Summary of findings	152
5.5.2	Prior Knowledge, VA and Realism	152
5.5.3	Observer Strategy	153
5.5.4	Object & Scene	154
5.5.5	Correlates of Visual Effort	155
5.6	Conclusions	155
6	Modelling Image-Wise Observer Sensitivity	157
6.1	Introduction	157
6.2	Related Work	158

6.2.1	Learning-based Human Perception Models	158
6.2.2	Unsupervised, Self-supervised & Transfer Learning	159
6.2.3	Semi-supervised Learning	160
6.2.4	Invariance & Equivariance	160
6.2.5	Transformation Equivariant Representations	161
6.3	Method	162
6.3.1	Overview	162
6.3.2	Distortions as Transformations	162
6.3.3	Perceptual Thresholds as Decision Boundaries	163
6.3.4	Psychometric Function Estimation	164
6.3.5	Transformation Equivariant Representation Learning	167
6.3.6	Perceptual Threshold Classifier (PTC)	170
6.4	Results	173
6.4.1	Perceptual Threshold Estimation	173
6.4.2	AET	176
6.4.3	Perceptual Threshold Learning	179
6.4.4	Application to Real Composite Images	182
6.5	Discussion	183
6.5.1	2AFC Study & JNDs	183
6.5.2	TER via AET	186
6.5.3	Approximation of Empirical JNDs	187
6.5.4	Towards a General Representation of Naturalness	187
6.5.5	Wider Applications	188
6.6	Conclusions	188
7	Applying Perceptual Models in Image Harmonisation	190
7.1	Introduction	190
7.2	Related Work	191
7.2.1	Image Compositing, Harmonisation and Deep Learning	191
7.2.2	Multi-task Learning, Feature Sharing & Attention	192
7.2.3	Attention instead of masks	193
7.3	Methodology	194
7.3.1	Overview	194
7.3.2	Approach	195
7.3.3	Test Dataset	195
7.3.4	Evaluation Procedure	197
7.3.5	Similarity Metrics	197
7.4	Results	197
7.5	Discussion	198
7.6	End-to-End Model: Methodology	200
7.6.1	Model Architectures	200
7.6.2	Optimization Details	203

7.6.3	Evaluation	203
7.7	End-to-End Model: Results	203
7.8	Discussion	206
7.9	Conclusions	209
8	Conclusion	211
8.1	Overview	211
8.2	Summary of Findings	212
8.2.1	Modelling Human Assessment of Image Composite Realism	212
8.2.2	Learning the Observer Function	215
8.2.3	Application to Harmonisation	216
8.2.4	Broader Impact	217
8.3	Limitations and Future Work	217
8.3.1	Data Collection and Perceptual Models	217
8.3.2	Model Optimisation, Scalability and Transformation Types	218
8.3.3	Application to Other Tasks & Alternative Approaches	219
8.4	Final Comments	219
	Bibliography	221
	A Publications	243

List of Tables

2.1	ITU five-grade quality and impairment rating scales.	39
4.1	The parameter ranges used to generate stimuli for the experiments.	86
4.2	Parameter values of logistic psychometric functions fit to the experimental data for each of the transformation features.	104
5.1	Experiment design	133
5.2	Comparison of fixations on targets across experimental conditions.	148
6.1	Cross-validation results	182
7.1	Means of similarity metrics for both techniques evaluated against ground truth.	198
7.2	Test metrics for all evaluated models, across the two datasets used in our experiments.	204
7.3	Average MSE on the iHarmony and COCO-Exp datasets for each of the evaluated models, grouped by area of harmonised object as a fraction of image size.	207

List of Figures

1.1	Examples of image composites with varying degrees of realism.	2
1.2	Example of visual difference between unharmonised and harmonised composites.	4
1.3	Idealised compositing pipeline.	5
1.4	Overview of logical structure and relationships between key chapters in thesis.	11
1.5	Overview of the methodology developed in this thesis.	11
2.1	Conceptual differences between feed-forward and feed-back models of human perception	16
2.2	Four Stages of Visual Processing	19
2.3	Empirical contrast sensitivity functions	22
2.4	Effect of decreasing relative contrast between background and target on the visibility of the target	22
2.5	Illustration of the effect of visual masking.	23
2.6	Overview of approaches to image quality assessment	26
2.7	Impact of JPEG compression on perceived quality.	26
2.8	Illustration of the error sensitivity framework for image quality assessment.	27
2.9	Example of the impact of Gaussian blur on the appearance of an image from the LIVE image quality database.	29
2.10	Pencil portraits of varying degrees of realism.	32
2.11	Contemporary photorealistic painting.	33
2.12	Examples of semantic and physical visual features undergoing violations. . .	34
2.13	Two prior assumptions in vision.	35
2.14	Single stimulus procedure	37
2.15	Double stimulus procedure	37
2.16	Double stimulus procedure	38
2.17	Average time required to compare N conditions under different experimental protocols.	39
2.18	Examples of standard test charts for visual acuity and normal colour vision.	42
2.19	Experimental stimuli from surface smoothness rating study.	47
2.20	Example of the output of the visual realism model by Lalonde and Efros (2007).	49

2.21	Example of statistical composite harmonisation	51
3.1	Neural network illustration	64
3.2	An Artificial Neuron	65
3.3	Empirical saliency maps	74
3.4	Architecture of the original DIH model.	76
3.5	Illustration of the spatial-separated convolution module attached to a U-Net style, encoder-decoder architecture. Image courtesy of Cun and Pun (2020)	77
3.6	An illustration of the architecture of DoveNet, including the generator with attention modules from Cun and Pun (2020) and two discriminator networks used in the adversarial training procedure.	78
4.1	Illustration of the synthetic composite generation process.	87
4.2	91
4.3	Examples of images used in the experiments.	93
4.4	High-level visualisation of the image dataset used in this study.	95
4.5	Illustration of some outliers from the image dataset.	96
4.6	Object sizes as fraction of image size.	96
4.7	An illustration of the experimental setup.	97
4.8	An illustration of the stimulus displayed to observers during each trial. . . .	99
4.9	Goodness of fit evaluation	103
4.10	Mean lapse rates	105
4.11	The exposure, contrast and CCT JNDs visualised for an image from the experimental dataset.	106
4.12	Visualisation of JNDs pt. 1	107
4.13	Visualisation of JNDs pt. 2	108
4.14	Visualisation of JNDs pt. 3	109
4.15	Images with highest average discrimination performance for stimuli with exposure transformations.	112
4.16	Images with lowest average discrimination performance for stimuli with exposure transformations.	113
4.17	Images with highest average discrimination performance for stimuli with contrast transformations.	114
4.18	Images with lowest average discrimination performance for stimuli with contrast transformations.	115
4.19	Images with highest average discrimination performance for stimuli with CCT transformations.	117
4.20	Images with lowest average discrimination performance for stimuli with CCT transformations.	118
5.1	Koch’s saliency framework.	123

5.2	Overview of Treisman’s Feature Integration Theory	125
5.3	Offsets applied to segmented objects in test images.	134
5.4	Experimental setup	136
5.5	Illustration of reference and test stimuli and associated interface presented to observers during the experiment.	137
5.6	Fixation heatmap example.	140
5.7	Realism responses averaged for feature offset values across image sets for exposure (left) and CCT (right).	143
5.8	Bootstrapped means/medians and their 95% confidence intervals for the evaluated metrics.	144
5.9	Bootstrapped comparisons of group mean/median differences for each of the evaluated metrics.	145
5.10	Area of interest similarity scores.	148
5.11	Comparison of joint fixation maps over reference and test stimuli, for each combination of factors.	150
5.12	Comparison of joint fixation maps over reference and test stimuli, for each combination of factors.	151
6.1	Illustration of just-noticeable differences for negative ($t-$) and positive ($t+$) local exposure transformations as a function of exposure shift.	159
6.2	Illustration of the 2AFC procedure used in the experiments.	165
6.3	VGG-16 architecture	168
6.4	Unsupervised AET architecture consisting of a VGG16-based convolutional autoencoder with weights shared across two inputs.	169
6.5	Illustration showing how key architectural elements from the AET network are incorporated in the PTC network. Red boxes indicate frozen weights. . .	171
6.6	Illustration of experimental results.	173
6.7	Visualisation of mean image-wise perceptual thresholds collected in the 2AFC study.	175
6.8	Mean MSE between ground truth and prediction for AET prediction errors for the validation dataset.	176
6.9	Illustration of AET output vs ground truth for a series of inputs.	177
6.10	Principal component visualisation of features extracted from images affected by local transformations of different magnitudes.	178
6.11	Predicted versus ground truth masks illustrating decision boundary pt. 1. .	180
6.12	Predicted versus ground truth masks illustrating decision boundary pt. 2. .	181
6.13	Performance on authentic composites.	184
6.14	Performance on authentic composites pt.2.	185
6.15	ROC curve illustrating the performance of the <i>PTC</i> when used for localisation of composited objects in the dataset from Xue et al. (2012). . .	186
6.16	Example of over-exposure and model correction.	187

7.1	System overview	194
7.2	Illustration of the preliminary two-stage evaluation of the standalone models.	195
7.3	Illustration of research methodology.	196
7.4	Dataset generation process	196
7.5	Similarity metric distributions for ground truth vs PTC masks.	198
7.6	Image-wise error differentials.	198
7.7	Examples of the DIH with ground truth masks over-compensating, and applying colour shifts to compensate a luminance transform, resulting in suboptimal output.	199
7.8	Comparison of harmonisation outputs from the evaluation.	201
7.9	Comparison of outputs from each model under evaluation for a range of images from the COCO-Exp dataset.	205
7.10	Comparison between the corrections applied by PTC+DIH, and the mask- based DIH-M models for multiple variants of the same image.	206
7.11	Examples of failure cases.	207
7.12	Examples of authentic composite images from Xue et al. (2012) processed using the <i>DIH</i> model.	208

Acronyms

- 2AFC** two-alternative forced choice. [73](#)
- AET** auto-encoding transformations. [167](#)
- AOIS** area of interest similarity. [139](#)
- CCT** Correlated colour temperature. [89](#)
- CG** computer graphics. [6](#)
- CNNs** Convolutional neural networks. [68](#)
- DFFO** duration of first fixation on object. [139](#)
- DIH** Deep Image Harmonisation Tsai et al. (2017). [193](#)
- DL** deep learning. [12](#)
- FCNs** fully convolutional networks. [72](#)
- GT** ground truth. [37](#)
- IOC** inter-observer consistency. [139](#)
- IQA** image quality assessment. [28](#)
- ITU** International Telecommunications Union. [38](#)
- JND** just-noticeable difference/distortion. [10](#)
- LPIPS** Learned Perceptual Image Patch Similarity. [197](#)
- ML** Machine learning. [54](#)
- MSE** mean squared error. [44](#)

PSNR peak signal-to-noise ratio. [44](#)

PTC Perceptual Threshold Classifier. [170](#)

PVQM Perceptual Visual Quality Metrics. [vi](#), [45](#)

TERs transformation-equivariant representations. [161](#)

TFFO time to first fixation on object. [139](#)

Glossary

digital image compositing the process of digitally assembling multiple images to make a final image, typically for print, motion pictures or screen display. It is the digital analogue of optical film compositing (Smith, 1995). [1](#)

fixation the maintaining of visual gaze on a particular location. [21](#)

saccade a transition of visual attention between two particular locations. [128](#)

visual realism the degree to which a viewed image of a scene produces the same response as the scene itself (Ferwerda, 2003; Fan et al., 2018). [1](#)

Chapter 1

Introduction

Digital image compositing combines visual elements from different sources to create the subjective impression of a single, coherent image (see Figure 1.1 for examples). Compositing is commonly performed manually, requiring considerable effort, as well as technical and artistic skill. The final quality, or visual realism, of image composites is commonly evaluated using subjective approaches, which are time-consuming to perform and impractical in the context of practical applications, such as film and visual effects, or mixed reality. Automation of both the assessment and improvement of the quality of image composites would allow for considerable improvements in terms of the cost and time efficiency. However, due to the complexity of human perception, modelling such complex phenomena is a challenging task. This thesis proposes techniques to address some of these issues, by modelling human perception of realism in the context of image compositing.

1.1 Motivation

1.1.1 The Importance of Visual Information

Digital images have become increasingly present in our lives. Visual information is paramount to human understanding of their environment and interaction with it, allowing for compact representation of complex phenomena. The rapid development of image acquisition and processing technologies, coupled with the proliferation of the Internet, have created an enormous amount of visual content and made it available worldwide. Methods of combining, transforming, and otherwise manipulating such images, for both aesthetic and practical purposes, have been developed and applied in a range of domains, from visual effects, through digital forensics to medicine. This process has been further accelerated by recent developments in the fields of machine learning, image understanding and synthesis, which have allowed for many complex image tasks to be addressed by learning nonlinear functions from example data, rather than being designed by hand.

Applications of image synthesis are wide-ranging, allowing for functional (Nie et al., 2017)



(a) Image by [Jörg Prieser](#) from [Pixabay](#)



(b) Image by [Stefan Keller](#) from [Pixabay](#)



(c) Image by [Josh Hild](#) from [Pexels](#)

Figure 1.1: Examples of image composites with varying degrees of realism.

and aesthetic (Wang et al., 2012) manipulation, domain transfer (Taigman, Polyak and Wolf, 2016) and generation of new image content (Nguyen et al., 2017). A wide range of approaches to image synthesis exist, including ground-up synthesis through computer graphics and rendering (Kajiya, 1986), statistically-based generative methods (Radford, Metz and Chintala, 2015), or recombination of existing image content, or elements thereof, into a seamless whole¹, known as compositing (Porter and Duff, 1984).

Arguably, image synthesis involves building an understanding of the causal factors behind visual content, their computational modelling and controlled reproduction or transformation. This process could involve transforming “geometry and physics into meaningful images” Glassner (2014), This is not unlike the field of natural language processing, which aims to accomplish similar goals with language-based content. In both scenarios, one of the key applications is to replace or aid the expensive human-in-the-loop, in order to allow automation of various processes, such as generation of new content, or comprehension and assessment of existing content. In both of these examples, humans are the ultimate recipients of such generated content, thus when designing such systems is important to ensure that their outputs align with human expectations. A classical example of such a subjective evaluation is the Turing Test (Turing, 1950, 2009), which evaluates how well a machine can emulate (or imitate) a human in a specific task. Turing’s work was a precursor to the development of machine learning and artificial intelligence in general, and formalised the philosophical underpinnings of machine intelligence. Many similarities can be found between the Turing Test and, in the scenario discussed here, subjective evaluation of the realism of a synthesised image.

1.1.2 The Fundamental Goals of Image Compositing

At a fundamental level, image compositing can be seen as the process of manipulating an image, or photograph, by inserting element(s) from another image, in such a way as to create a result which appears realistic and plausible to the average observer (see Figure 1.2 for an example). Most photographs of natural scenes can be seen as perfect image composites of the objects and scenes featured in them (Xue et al., 2012). Effectively, they capture the result of a complex rendering function, which considers the interactions of light and all parts of the scene. However, in the practical compositing scenario, when a region of an image is inserted into another image, it is very difficult, and often practically impossible, to access and re-apply the original rendering function to the inserted object. This results in perceptible differences between the distributions of certain image features, which can negatively impact subjective realism judgments, particularly if these differences occur across multiple features.

The set of natural images, such as those including objects and scenes humans encounter daily, is highly diverse. Variations within this set occur along many dimensions, including

¹A seamless whole can be seen as the ultimate goal of compositing, however ‘bad composites’ have become a fine art in recent years, see <https://youtu.be/6w6FV8P7HXg?t=41s> for example



Figure 1.2: Example of visual difference between an a) unharmonised composite and b) it's harmonised version, created by applying local colour correction. Image courtesy of Wright (2013b).

shape, size, illumination, position, colour or orientation, to name just a few. Respectively, the set of image composites – arbitrary combinations of elements from different source images – contains even more potential variety. This includes combinations of visual elements that would be physically impossible in the real world. Coupled with the complexity of the human visual system, this makes exhaustive modelling of human perception in this domain a challenging task, as it is infeasible to measure human response to distortions along all these dimensions of variation.

Interestingly, when a compositing artist performs the task manually, they commonly focus on adjusting a relatively small set of attributes, with the goal of matching the appearance of the foreground object to that of the background scene, often achieving perceptually realistic results (Wright, 2013b). This highlights two key points:

- the key focus of image compositing is minimising appearance-based differences between the object and scene
- many distortions common to image composites can be both plausibly corrected and introduced using approximate transformations

The compositing problem can thus be seen as an optimisation procedure based on applying transformations in order to minimise the perceptual difference between distributions of relevant features of the object and scene. The fact that composites *can* be made both realistic and unrealistic implies the existence of thresholds below which certain image-based differences are no longer perceptible to human observers. However, because of the multiple dimensions along which the elements of a composite images may differ, it is not straightforward to determine the most relevant features to correct, or the amount of correction required. Consequently, the process of compositing is often referred to as an art, as no universal solutions exist. Compositing artists commonly rely on experience and intuition regarding what looks “good enough” and, in some way, implicitly learn generalised perceptual thresholds of audiences. Understanding and modelling this process would allow for new methods for automatic assessment and improvement of image

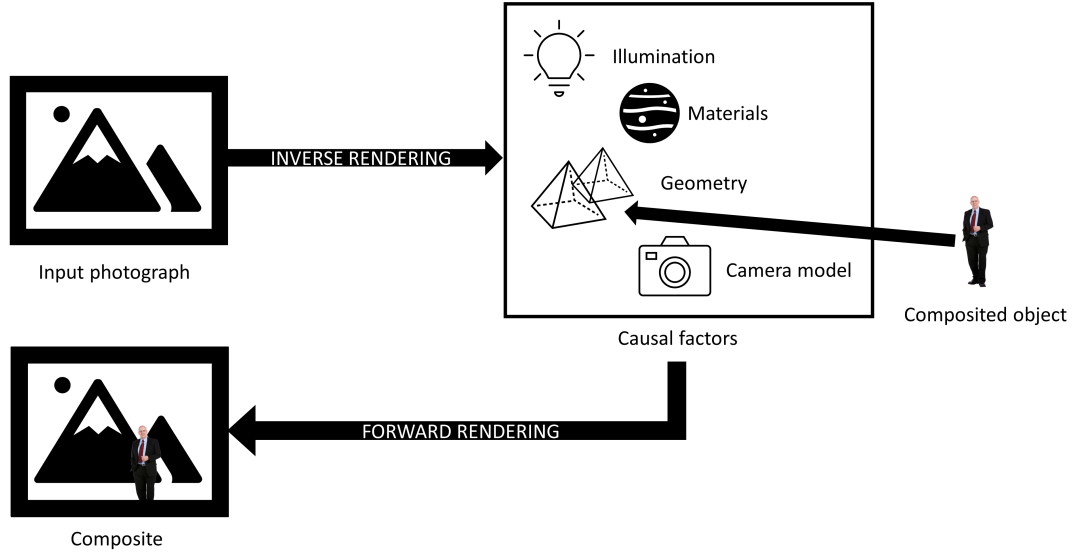


Figure 1.3: A theoretical idealised compositing pipeline, consisting of an input image, inverse rendering (decomposition of the photograph into geometry, illumination, material properties and a camera model), the insertion of a new element into the decomposed scene, followed by a forward rendering step. The forward rendering step captures an image of the composited scene, including the newly inserted object, whose appearance is affected by the scene properties (e.g. the illumination and camera model)

composites, based on human perception.

1.1.3 Perceptual Challenges of Image Compositing

In theory, a perfect image composite can be created through *inverse rendering* (Marschner and Greenberg, 1998) – the process of disambiguation and disentanglement of the causal factors behind a photograph – the scene geometry, illumination, material properties and camera model (see Figure 1.3 for a conceptual illustration). Once these properties are disentangled, foreign objects can easily be inserted and affected by the same appearance model as the original scene at the time of capture. This is followed by conventional forward rendering, which generates the composited image. In practice, due to the ill-posed nature of the inverse rendering problem, its results are commonly only approximated, either heuristically, by compositing artists, or through use of image-statistical approaches.

As such, some key questions, inevitably pondered by compositing artists, are raised when considering this approximation process, namely:

- Are all appearance-based properties equally important to achieve subjective realism?
- If not, which properties should one focus on to achieve a ‘good enough’ result, without wasting effort?
- How much do human observers vary in their subjective perceptions of these

properties?

- Do observers use common visual strategies in assessing the quality, or realism, of image composites ?
- How does scene content affect these strategies?

The success of both special photographic effects, and visual effects in the film industry over the last 50 years (Prince, 2010), suggests that many compositing artists have learned to build intuitive models of audience expectations and visual sensitivity (Prince, 2011). Accordingly, compositing artists have, at least implicitly, come up with practical answers for the above questions. At least to the degree that their creations have been considered plausible, realistic and immersive by audiences. Equally, the expectations of cinema audiences have grown along with their experience of the medium and content (Kane, 2011). A good example is how special effects in a science fiction films from the 1950s appear dated, and often comically unrealistic, compared to modern digital visual effects.

1.1.4 Practical Challenges of Image Compositing

Typically, compositing is carried out in a manual or semi-automatic manner, using a variety of software tools, such as Adobe Photoshop, After Effects or Nuke 3D. The process involves the use of a variety of image processing operations, which aim to minimise noticeable differences in certain properties of source images, such as brightness, contrast or saturation. The type of image processing operations used, and their parameter settings depend vastly on the source images and the conditions under which they were captured, or generated. For example, when combining computer graphics with live action content, it is possible to manipulate additional properties of the rendered computer graphics (CG) content, such as illumination, texture, albedo and so forth.

The sheer number of possible parameters and a near infinite set of combinations of image elements makes compositing a complex problem, often solved through painstaking incremental manual corrections. While automatic compositing methods do exist, for example in mixed reality headsets, they tend to produce composites of inferior quality, compared to manually-created ones (Klein and Murray, 2009). The process of quality evaluation is often performed in a manual, subjective and often informal manner (trial-by-peers, or a focus group approach), and thus is time-consuming and costly. Additionally, the manner in which humans perceive inconsistencies in images is nonlinear and difficult to reconcile with how images are stored in digital systems. This complicates any reasonable optimisation of compositing efforts guided by minimum audience expectations or any consistent indication of the aspects of the composites that need urgent improvement. The above factors make it difficult to successfully automate compositing and its subsequent quality evaluation. This, in turn, makes the process prohibitively time-consuming and steepens the learning curve for novice users.

In order to tackle these issues without sacrificing creative input, automatic prediction of composite quality should be capable of guiding the artist towards a more realistic final result, while still allowing for creative freedom. Accordingly, a detailed understanding of human perception of distortions common in compositing, the contributing parameters and their mapping to subsequent quality judgements would allow to rank and localise regions of poor quality. A system able to automatically predict subjective quality ratings could then also be automated to correct them more efficiently. Aside from iterative assistance in the creative process, this would allow to significantly speed up a range of quality control and compositing tasks in general. This would have positive effects for both novice and professional users, by allowing for more time to focus on the artistic and creative aspects, while speeding up the coarse matching and achievement of a baseline satisfactory quality. The work presented in this thesis focuses on developing such perceptually-inspired, localised composite quality assessment metrics able to inform both automatic and semi-automatic compositing methodologies.

1.2 Application Domain

The contributions made in this thesis are applicable in a range of domains of image synthesis and evaluation. The primary application area and chief source of inspiration for this work is in visual effects and compositing - the synthesis, transformation and generation of new image and video content by combining natural images with other synthesised or captured image content. This is because the techniques developed in this thesis attempt to automate parts of the compositing process in order to aid artists and allow them to focus on the creative, rather than technical, aspects of their work. Aside from this key application area, other relevant applications are summarised and discussed below:

1.2.1 Visual Effects

Production of visual effects for film and TV is an expensive process, requiring significant manual effort (Wright, 2013a). In major motion picture productions, visual effects production is often allocated a third of the total budget (Curtin and Vanderhoef, 2015). Automation of various aspects of the compositing pipeline, such as colour correction or harmonisation, could allow for increased efficiency and reduction in costs, as well as allowing artists to focus on creative aspects of the process. The models proposed in this thesis can be incorporated as plugins into such pipelines. Additionally, the framework for learning to detect and correct perceptually-based image transformations developed in this thesis can be applied to tasks other than composite harmonisation. Like in other learning-based scenarios, this can be accomplished through appropriate selection/generation of training data.

1.2.2 Subjective & Perceptual Metrics

Digital image data has become extremely commonplace. Many common tasks involving images rely on various metrics to index, retrieve and filter image data in large databases of images (Cui, Wen and Tang, 2008). As most image data is ultimately targeted at human observers, such metrics should correlate with human perception and preferences. For example, the outputs of many algorithms operating in the image domain, for example image compression (Patel, Appalaraju and Manmatha, 2019), superresolution (Wang et al., 2017) or colourisation (Zhang, Isola and Efros, 2016) are often evaluated with respect to human perception. Accordingly, various perceptual image metrics have found extensive use in practical applications, for example as objective metrics to minimise in various optimisation-based tasks (Zhang et al., 2018b). Whether it’s image quality, aesthetics, realism, saliency or visual similarity, automatic measurement of subjective properties in image data is significantly improving how increasingly vast datasets of image data are organised, explored and managed (Gordo et al., 2016).

1.2.3 Image Forensics

While manipulation of images is not novel in and of itself, recent developments in deep learning-based image synthesis techniques (Goodfellow et al., 2014; Reed et al., 2016; Zhang et al., 2018a; Wang et al., 2018a; Park et al., 2019) have made image manipulation techniques much more powerful and accessible to a wider potential audience. Consequently, there is increased need for tools to detect such manipulation (Ferreira et al., 2020; Yang et al., 2020). The models proposed in this thesis contribute to this body of work, as they are designed to detect perceptible image manipulations, or distortions, not present in natural images. Provided appropriate exemplar data, this methodology could be extended to detection of other types of image manipulation artefacts, not necessarily only ones perceptible to humans.

1.2.4 Human Observer Modelling

The techniques developed in this thesis serve as an approach to human observer modelling (Geisler, 2011). In many complex systems, such as motion picture production, manufacturing, quality assurance, visual inspection by humans is a common and often expensive requirement. Predicting human visual performance in such practical scenarios often allows for significant gains in efficiency (Gai and Curry, 1976). Aside from detailing the visual strategy observers adopt in the task of composite realism assessment, this work proposes general-purpose approaches to learning-based approximation of human performance in particular visual tasks. This is based on a combination of empirical perceptual modelling and approximation of the resulting perceptual functions using gradient-based optimisation techniques. This effectively allows for generalisation of an empirical model to novel image content. As such, complex subjective visual tasks can be modelled using this approach, allowing to address visual tasks otherwise requiring a

human in the loop (Cranor, 2008).

1.3 Research Aim

The overall aim of this thesis is to produce perceptually-driven systems to facilitate computational subjective quality analysis of image composites and guide the subsequent manual or automated improvement of their quality.

1.4 Research Objectives

The following questions are answered in order to achieve the aim:

1. Define subjective quality and visual realism be in the context of digital image composites.
2. Measure whether controlled changes in features of image composites predict subjective quality/realism judgements.
3. Model how changes in different composite image features affect subjective quality/realism judgements.
4. Assess whether subjective quality judgements are affected by VA and correlated with spatial image regions.
5. Generalise psychometric models to map localised image features to predictions of subjective quality judgements.
6. Evaluate if the generalised models provide dependable quality assessment and improvement on related tasks.

1.5 Methodology

To address the aim, the following methodology is adopted. First, visual realism is defined in the context of image composites, as a function of perceptible, local, and object-based image differences. This definition is then evaluated in an empirical study measuring group JNDs for *synthetic composites* - images with local image transformations approximating common distortions found in authentic image composites. This is performed for a large sample of human observers. Psychometric models are then fit to the empirical data. Additionally, visual attention allocation and impact of prior knowledge are evaluated in a separate experiment, in order to assess the impact of scene content on the proposed psychometric models. In order to generalise these psychometric models to a wider set of input stimuli, learning-based techniques are adopted to approximate the empirical functions. This is achieved by self-supervised learning of a TER by learning to predict local transformation parameters, followed by fine-tuning of this model using the empirical

psychometric models from earlier experiments. Finally, the trained models are evaluated in a related downstream task - image harmonisation. This methodology is illustrated in Figure 1.5.

1.6 Thesis Structure

Chapter 2 presents the background on compositing, the human visual system, approaches to its modelling and definitions of visual realism, as well as related work in image quality assessment, visual realism prediction and automatic image compositing and harmonisation

Chapter 3 discusses the background on machine learning, gradient-based optimisation and deep convolutional neural networks, presenting an overview of key concepts and applications relevant to the topic of this thesis.

Chapter 4 proposes a methodology for modelling human perception of visual realism as a function of transformation magnitude, and presents experimental results and psychometric just-noticeable difference models for several common composite distortions represented as local image transformations.

Chapter 5 investigates the spatial and temporal allocation of visual attention in subjective realism assessment, as well as evaluating the impact of prior knowledge and transformation type on visual attention and subjective realism ratings.

Chapter 6 presents a methodology for generalisation of the proposed empirical just-noticeable difference/distortion (JND) models using deep convolutional neural networks and representation learning.

Chapter 7 applies the proposed models to a related task - no reference image composite harmonisation - and presents an evaluation, comparing the model against baselines.

Chapter 8 summarises and discusses the findings from previous chapters, detailing broader application scenarios, limitations and directions for future work.

The overall logical structure of this thesis is also illustrated in Figure 1.4

1.7 Contributions

The chief contribution of this thesis is the proposal of an end-to-end methodology for measurement, modeling and reproduction of human visual sensitivity to common image-based compositing artefacts. A number of other contributions are made in the process of achieving this goal, namely:

- A novel method for measuring visual realism as a function of distortion visibility and a set of resulting JND models (Chapter 4)

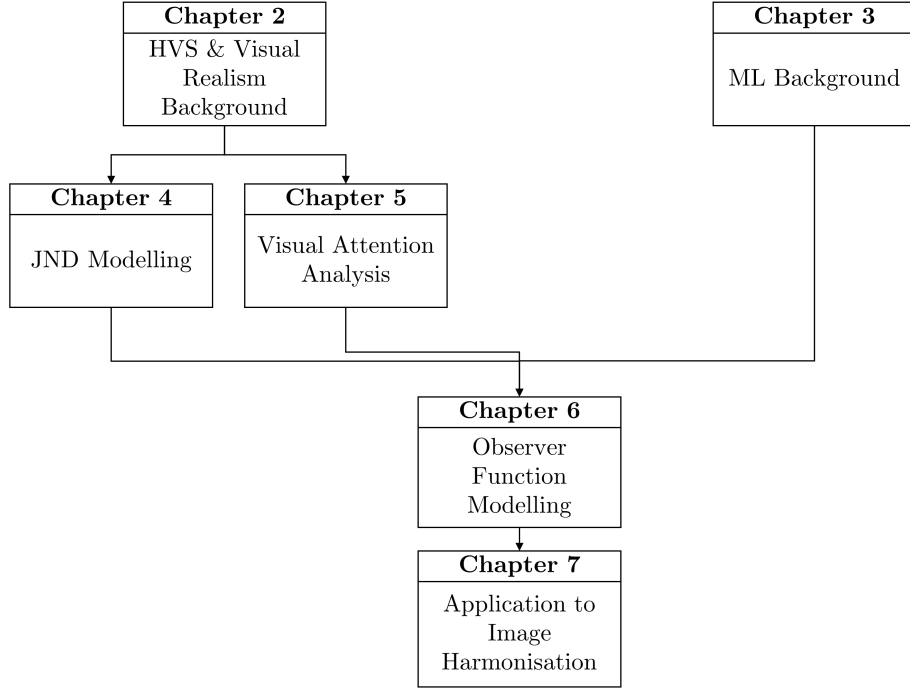


Figure 1.4: Overview of logical structure and relationships between key chapters in thesis. Chapters 1 and 8 excluded for clarity.

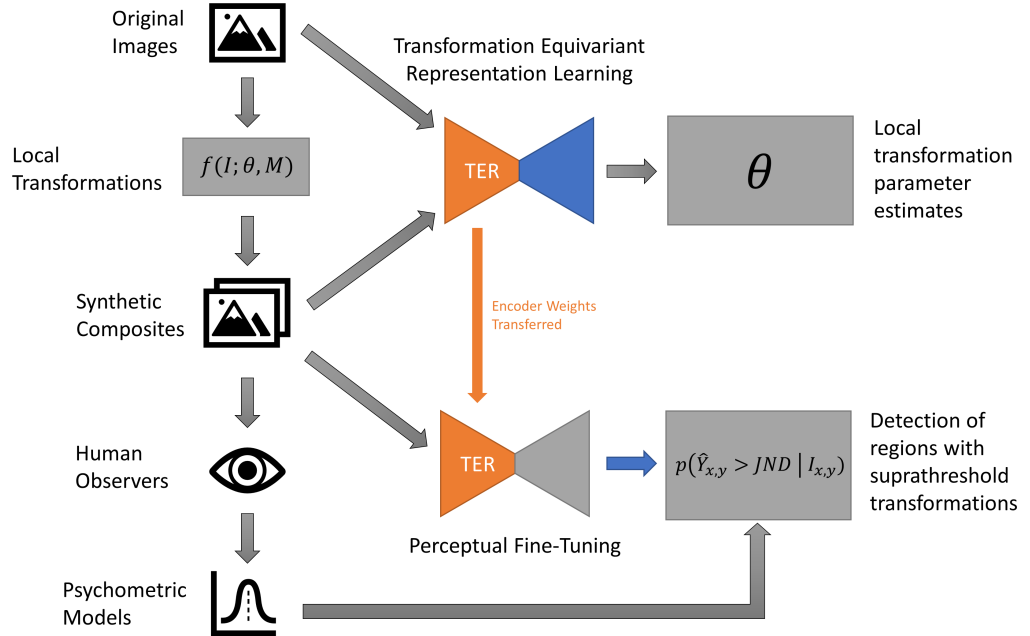


Figure 1.5: Overview of the methodology developed in this thesis.

- A study into how visual attention is deployed during subjective image composite assessment, illustrating the impact of prior knowledge and distortion type (Chapter 5)
- A methodology for generalisation of JND-based models to unseen images using deep learning (DL) techniques (Chapter 6)
- An application of generalised JND-based models to an image harmonisation problem (Chapter 7)
- A state-of-the-art image composite harmonisation model requiring no input masks (Chapter 7)

The following papers have been published as part of this work:

1. Dolhasz, A., Williams, I. and Frutos-Pascual, M., 2016. Measuring Observer Response to Object-Scene Disparity in Composites. *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE, pp.13–18
2. Dolhasz, A., Frutos-Pascual, M. and Williams, I., 2017. Composite Realism: Effects of Object Knowledge and Mismatched Feature Type on Observer Gaze and Subjective Quality. *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE, pp.9–14
3. Dolhasz, A., Harvey, C. and Williams, I., 2020. Learning to Observe: Approximating Human Perceptual Thresholds for Detection of Suprathreshold Image Transformations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp.4797–4807
4. Dolhasz., A., Harvey., C. and Williams., I., 2020. Towards unsupervised image harmonisation. *Proceedings of the 15th international joint conference on computer vision, imaging and computer graphics theory and applications - volume 5: Visapp.*. INSTICC, SciTePress, pp.574–581. Available from: <http://doi.org/10.5220/0009354705740581>

In addition, the following workshop presentations and invited talks were given:

1. Williams, I., Dolhasz, A., & Monnayer, N. (2015, September). Measuring Perception of Realism in Mixed and Augmented Reality. *In 2015 IEEE International Symposium on Mixed and Augmented Reality Workshops (pp. 24-24)*. IEEE.
2. Dolhasz, A., 2017. Towards a more human machine perception of realism in mixed reality. *Whitehead Lecture Series, Goldsmiths University of London*
3. Dolhasz, A. 2020. Perceptually-based Detection of Local Exposure Transformations. *Warwick University Visualisation Group (virtual)*.

Chapter 2

Understanding Perceptual Image Quality & Visual Realism

2.1 Introduction

Humans perceive the visual world around them with little conscious effort, able to extract and process a wide range of complex properties of their environment in a fraction of a second, relying on limited information. From the two-dimensional array of light intensities falling on their retinæ, they are able to effortlessly infer three-dimensional features of objects in their environment, such as shape, position, colour or illumination. This allows them to navigate spaces, interact with objects and perform a wide variety of other tasks relying on visual information. Despite this, humans are also capable of impressive neglect of visual information, for example, failing to notice large changes in visual scenes (Simons and Levin, 1997), or being fooled by visual illusions (Eagleman, 2001). This variability inherent in the human visual system, further influenced by task, context, attention and personal experience, while allowing for the visual system to perform under a range of conditions, makes accurate prediction of subjective perceptual properties a challenging task.

Computer systems, on the other hand, while capable of storing and transforming image data perfectly, cannot (yet) intrinsically ‘perceive’ the content of images the same way humans can.¹ With recent advances in artificial intelligence, machine learning and pattern recognition, particularly deep neural networks (discussed in Chapter 3), combined with increases in computational power, these operational-level differences between human and computer perception and related task performance are starting to diminish. For instance, human performance has already been matched by computers in particular tasks, such as image recognition, house number classification in street view images, or playing computer games (Eckersley and Nasser, 2017).

¹However, due to the pace of research in this area, the statement above is becoming less accurate by the day.

The problem of image quality assessment, and similarly the assessment of many subjective properties of digital images, combines both problem paradigms illustrated above:

1. understanding of the human visual system and measurement of its performance of subjective tasks
2. computational modelling of these processes, such that they can be applied to digital images in practical scenarios.

Humans are known to assess image quality based on properties, or distortions, that are noticeable to them, rather than objectively quantifiable from the raw image data (Yang et al., 2005). This is both due to the varying sensitivity of the human visual system to different types of distortions and patterns, as well as the uneven distribution of visual attention, often attracted by salient or task-relevant regions. These properties have been broadly exploited, particularly in the domain of broadcast, where subjective quality testing has been successfully used to evaluate compression and transmission artifacts, aiming to minimise the amount of data required to represent a signal, whilst keeping the subjective quality at an acceptable level for the average audience member (Wang and Li, 2010). Conversely, computer systems are better suited to judging quality by reference, e.g. comparing a digital image before and after compression and calculating a measure of pixel-wise differences between the two (Sheikh, Sabir and Bovik, 2006). However, extensive training by example, or complex computational models of the human visual system are required in order for a computer system to approximate subjective human judgments, particularly when no reference image is provided to compare against. Consequently, in order to address the issue of subjective image quality in the domain of image composites, it is necessary to first review relevant properties of human visual perception, methods for measuring subjective responses to visual stimuli and the unique properties of image composites that set them apart from natural images.

This chapter begins by presenting relevant background on human visual perception and image quality assessment. This is followed by a literature review of methods for measurement, modelling and prediction of image quality. Background on digital image composites and the properties that set them apart from natural images is then presented. Next, the concept of visual realism is discussed in the context of its definitions in prior work. Practical relationships to image quality are then developed via a review of relevant literature. Finally, existing methods for measurement and modelling of visual realism are reviewed. Building on these ideas, the concept of *visual realism* is then generalised as a special case of image quality.

2.1.1 Narrative Structure

This chapter discusses a broad range of topics, crucial to understanding and discussing the topic of visual realism. This section provides a summary and overview of how the different topics discussed in this chapter. First, relevant background on human visual

perception is presented in Section 2.2. This leads to a discussion of modelling of image properties, such as image quality. The concept of image quality is then introduced in Section 2.3, and its relationship to visual realism is established in Section 2.4. Once both concepts are defined and discussed, relevant practical frameworks for measurement and modelling of subjective quality and realism are reviewed in Section 2.5, including grading scales, experimental designs, stimulus presentation and observer selection, while Section 2.6 introduces objective approaches to the same problems. Related methods for measuring and modelling visual realism are discussed in Section 2.7. Finally, Section 2.8 summarises the information introduced in this chapter.

2.2 Background on Human Visual Perception

Visual perception is the process of acquiring knowledge about environmental objects and events by extracting information from the light they emit or reflect.

Palmer (1999, p. 5)

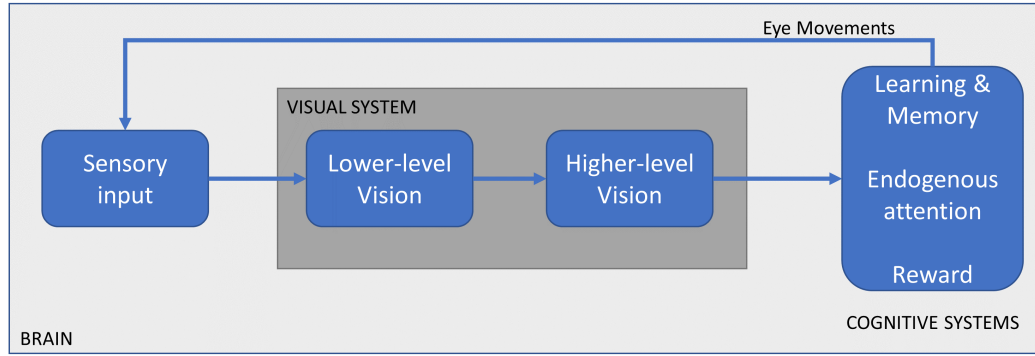
Humans are able to reliably recover properties of the three-dimensional world around them using only a two-dimensional retinal input. Despite the fact that a single 2-D retinal image can be generated by infinitely many three-dimensional arrangements of objects, humans are able to accurately and veridically recover many 3-D properties of the world, under a range of observation conditions. For example, despite the fact that the 2-D retinal images of objects change shape and size depending on viewing angle and position of the observer, the properties of the 3-D percept of the object remain consistent. Conversely, changes in certain properties of a visual stimulus, such as angle of illumination or shadow direction, can go unnoticed or require significant effort to detect (Ostrovsky, Cavanagh and Sinha, 2001). This suggests that an understanding of relevant aspects of human visual perception is critical to modelling assessment of subjective image properties, error detection and visual evaluation. This section discusses visual perception, the fundamental characteristics of the human visual system, and the application of this knowledge to the development of practical models of human perception.

2.2.1 Direct vs Inverse Problems

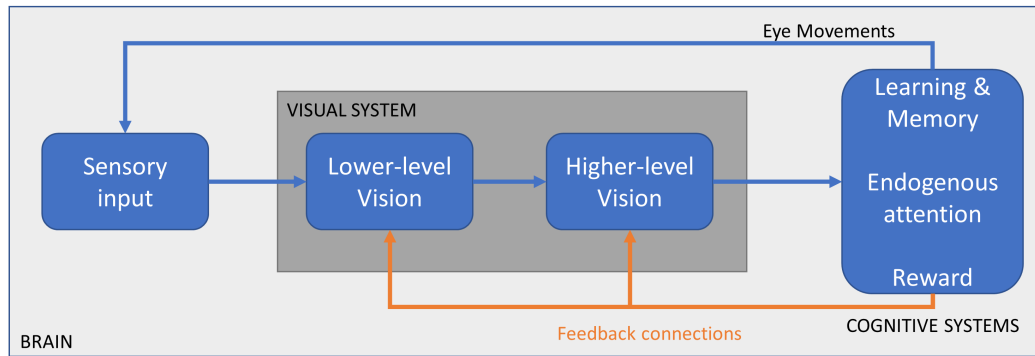
The human visual system (HVS) and visual perception have attracted a significant amount of research interest for centuries (Helmholtz, 1856). Throughout this time, two schools of thought in visual perception have emerged. The key distinction between them is the representation of the problem of extracting 3-D information from a 2-D image as either a direct or inverse problem (see Figure 2.1 for an illustration).

Perception as a direct problem

The direct perspective posits that perception is a bottom-up, data-driven process where the stimulus is transformed into increasingly complex, or high-level, representations between



(a) Feed-forward / Bottom-up / Direct Model of Human Perception



(b) Feedback / Top-down / Inverse Model of Human Perception

Figure 2.1: Illustration of the conceptual differences between a) feed-forward and b) feed-back models of human perception. Reproduced from Emberson (2019)

the retina and the visual cortex. This paradigm can be likened to the process of successive extraction of features from the retinal image. This school of thought builds largely on the work of Gustav Fechner and his seminal “Elements of Psychophysics” (Fechner, 1966). Fechner claimed that percepts are results of sensory inputs, and understanding them requires measuring responses to physical stimuli in a controlled manner. Visual perception as a direct problem was formally outlined by Gibson (1966) who suggested that no higher cognitive input was necessary for perception to function. Gibson’s theory relied largely on observer motion to extract useful information.

Perception as an inverse problem

Visual perception as an inverse problem considers a top-down, hypothesis- or expectation-driven perspective. The inverse problem paradigm models perceptual processes through the use of *a priori* information to constrain the ambiguity introduced by the projection of visual information from 3-D to 2-D at the interface between the outer world and the visual system (i.e. the retina). Bayesian methods are often employed in order to achieve this. This school of thought dates back to the work of Von Helmholtz (1867) who suggested that visual perceptions are unconscious inferences from sensory input and past knowledge of the world. His work was extended by Gregory (1970) who described perception as

a visual hypothesis, giving predictions about unseen properties of objects in the world. Gestalt studies of perception (Wagemans et al., 2012) also relied on the inverse problem definition while adopting a more holistic, qualitative approach, compared to Helmholtzians. One of the central principles of gestalt perception was the law of prägnantz (Good Figure, Law of Simplicity), stating that ambiguous or complex images would be perceived and interpreted in their simplest form possible. Despite some criticisms, gestalt principles became the foundation of recent work in computational shape perception, proposed by Pizlo (2014). In his seminal work, Marr (1982) also emphasised the role of top-down constraints, suggesting that understanding *why* certain information processing occurs is as important as understanding *how* it occurs.

Perception as predictive processing

To date, both the direct and indirect paradigms have resulted in significant advances in the understanding of human vision. The current consensus is that both top-down and bottom-up processing are required to explain processes in vision (Poggio, Torre and Koch, 1987; Li and Pizlo, 2011). In recent years, the **predictive processing framework** has been referenced as a potential route for reconciliation of the two schools of thought and proposed as a unified theory of cognition (Clark, 2013). The predictive processing framework proposes that perception and cognition rely on a bidirectional interaction of top-down predictions and bottom-up sensory signals. Due to the inherent ambiguity of real-world visual stimuli, the role of top-down predictions is to provide constraints based on learning, prior knowledge and expectations, which in turn help disambiguate sensory signals. When the proposed prediction cannot fully explain the stimulus, the residual prediction error is propagated forward, triggering a learning process that leads to revision of hypotheses (learning) (Lupyan and Clark, 2015). This “perception as hypothesis testing” school of thought, has recently been formalised and detailed by Hohwy (2013).

This view of visual perception emphasises the importance of both bottom-up modelling of the visual stimulus and its features, as well as the top-down aspects of the subjective viewing experience. This allows for the influence of task, experience, attention and other subjective traits to be modelled independent of the stimulus. Under this paradigm, perception can be modelled as a combination of direct measurement and inference constrained by some ecological, or learning-based priors.

2.2.2 Computational Vision

To date, perhaps the most comprehensive and pragmatic attempt to theorise and model human visual perception has been proposed by Marr (1982). His discussion of vision adopted a computational paradigm to describe the HVS as a hierarchical, modular set of information processing systems. He argued that due to its inherent complexity, in order to focus on and solve specific problems, such an information processing system should be analysed at *three separate levels* of abstraction:

- **computational** - understanding what the goal of the input-output computation is, the reasons for its appropriateness and the logic by which the computation can be carried out
- **algorithmic** - recognising how such a computational theory can be implemented, i.e. how are the inputs and outputs represented and what algorithm is used for the transformation
- **implementational** - understanding how the representation and algorithm can be implemented and realised, both physically and biologically.

These levels of analysis, the emphasis on understanding the *goal* of computations, as well as studying their algorithmic implementation, are perhaps Marr's greatest contributions, emphasising the importance of both top-down memory and cognitive processes, as well as bottom-up processing of information and its representations. This approach to studying visual cognition across multiple-levels of analysis was fundamental to the development of predictive processing theories (Bubic, Von Cramon and Schubotz, 2010), early connectionist ideas (Fodor, Pylyshyn et al., 1988; Smolensky, 1988), and Bayesian approaches (Knill and Richards, 1996) to modelling human perception.

2.2.3 Stages of Visual Processing

Drawing on Marr's framework, at the *computational* level of analysis the overall goal of perception is constructing an internal model of the physical environment within our field of view in order to be able to efficiently interact with it. This involves a transformation from a two-dimensional retinal input to an output: a three-dimensional description of the local environment. Based on this paradigm, the perceptual process can be broken down into 4 stages, according to Palmer (1999) and based on Marr's analytical framework. Each of the stages described by Palmer is characterised by distinct input and output representations and the respective processes required to transform the input into the output representation.

1. The Retinal Image

Visual perception begins with the acquisition of a retinal image (the proximal stimulus) from a projection of the external environment (the distal stimulus). This retinal representation is often approximated as a homogeneous 2-D array of receptors I , each of which occupying a discrete location, represented by its (x, y) coordinates in retinal image space and sampling the intensity of the light incident on it.

2. The Image-Based Stage

Once the proximal stimulus is sampled at the retina, the first stage of visual processing derives progressively higher-level representations of the array of image intensities, such as:

- extraction of primitives such as local edges, lines or corners from the intensity image,

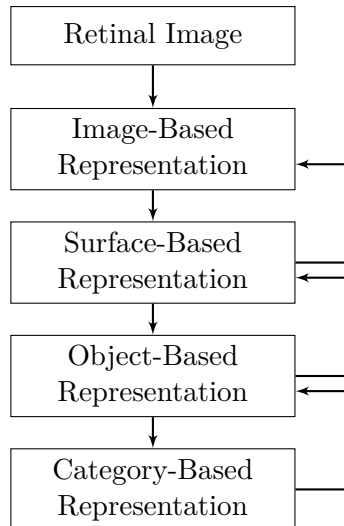


Figure 2.2: Visual perception represented as four stages of visual processing, from the retinal image to categorised objects. Note that the arrows indicate both bottom-up and top-down flow of information. Reproduced from Palmer (1999).

- defining 2-D regions in the image using edge information
- grouping and organisation of primitives into larger structures, based on their properties,

Marr referred to the results of this image-based stage as the *primal sketch*. Specifically, he referred to the result of the initial detection and localisation of these primitives as the *raw primal sketch*, and the result of the grouping process as the *full primal sketch*. While some disagreement as to the specific processes within this stage can be found in the literature, the consensus is that the output representation of this stage is:

- spatially relative to the retinal coordinate system,
- inherently two-dimensional and,
- comprised of image-based primitives, which describe the structure of the proximal, rather than distal stimulus.

3. The Surface-Based Stage

This stage is concerned with recovering properties of visible surfaces from the image-based representation. It differs from the former stage in that it represents distal stimuli in terms of spatial layout of their surfaces in three dimensions, from a viewer-centred reference frame, rather than focusing on 2-D image features in a retina-centred reference frame. The surface-based representation consists of a number of surface primitives – local patches of 2-D surfaces at a particular distance and slant from the viewer. Initially proposed by Gibson (1950), this stage recognised the role of depth and surface orientation in recovering 3-D information. Subsequent work

by Marr (1982) and Barrow and Tenenbaum (1978) confirmed the importance of a surface-based representation, particularly its encoding of properties of surfaces in the external world, rather than those of the retinal image. They also suggested that the process of extracting such surface representations makes use of particular image-based properties such as *texture*, *shading*, *motion* and *stereoscopy*. Subsequent successful application of these processes in the field of computer vision reinforced the value of a surface-based representation (Longuet-Higgins, Prazdny et al., 1980; Longuet-Higgins, 1981).

4. The Object-Based Stage

This stage of perception is concerned with integrating the incomplete 2.5-D information from the surface-based representation into a true 3-D representation, which includes some unseen, occluded surfaces in addition to those from the previous stage. Thus, this stage is represented with volumetric primitives within a 3-D space and an object-based reference frame. This stage of perception clearly illustrates the importance of top-down assumptions and pre-existing constraints, as the information presented to the observer is incomplete, yet the representation derived from this information appears to be veridical. To date, there is no clear consensus regarding how exactly this representation is arrived at from the surface-based representation. Perhaps the most fundamental paradigm change has recently been proposed by Pizlo (2014) who argues the importance, if not superiority, of *a priori* constraints over bottom-up visual information in the context of object shape perception, providing compelling evidence.

Marr’s modular approach to the analysis of vision provides both a computational framework for the analysis of visual processes, as well as computational models for several visual processes, such as edge extraction, grouping or structure from motion. Many modern computer vision techniques also rely on the hierarchical approach to modelling vision, particularly Convolutional Neural Networks, discussed in Section 3.3.5.

2.2.4 Scene Perception & Organisation

In order to perceive the complex world around us, mere 3-D object perception is not enough, as these objects are often combined and organised into more complex scenes. When humans perceive real-world scenes, they are able to internalise much more than simply a listing of objects present in the scene. In fact, it is well-known that context - the semantic and spatial relationships between these objects - is crucial to the correct identification of scenes (Biederman, 1972). Humans are extremely efficient at the task of scene classification and are able to encode properties such as relative size, positions, semantic probabilities and interactions between objects in a very short time, as evidenced by Biederman, Mezzanotte and Rabinowitz (1982). In fact, Potter (1976) noted that understanding of an average scene requires around 100ms, while a memory representation requires another 300ms. However, what is captured by the visual system is likely not a

detailed ‘photographic’ representation, but rather a coarse, abstract one, referred to as the ‘gist’ of a scene (Friedman, 1979; Intraub, 2002). Oliva (2005) claims that global image features are responsible for this time efficiency, ruling out a full, bottom-up process due to ecological implausibility and evidence that a single fixation can be enough to correctly classify a scene (Intraub, 2002).

2.2.5 Colour Vision

Human perception of colour has been theorised on two distinct levels. The trichromatic theory of colour vision proposed by Thomas Young (Young, 1802) poses colour vision as a function of three different receptors in the eye, each sensitive to a specific colour. This theory was extended and formalised by Helmholtz (1856) who identified the receptors as cones sensitive to different wavelengths of light - long (L), medium (M) and short (S) - corresponding to primary colours. Conversely, the opponent-process theory of colour vision, proposed by Ewald Hering (Hurvich and Jameson, 1957), describes colour vision as a function of the brain and its interpretation of the signals from the eye. It builds on the Young-Helmholtz theory and describes colour perception as an opponent process, relying on excitatory and inhibitory combinations of signals from the L, M and S cones. These combinations form three colour-opponent channels, two chromatic (blue-yellow and red-green) and one achromatic (white-black).

2.2.6 Contrast Sensitivity

The HVS is less sensitive to absolute changes in stimulus, compared to relative ones. Consequently, it is the differences in colour and luminance that make it possible for the HVS to distinguish patterns and objects in the real world. The ability of the HVS to detect details of a visual scene hinges on both the relative size and contrast of those details. Campbell and Robson (1968) evidenced this by measuring the human contrast sensitivity function (CSF), describing the relationship between the spatial frequency of sinusoidal gratings and their visibility to observers. The CSF is important for quantifying the performance of the visual system (Van Nes and Bouman, 1967), as it provides a measure of the frequency response of the HVS. Practically, its employment allows for local weighting of visual stimuli based on their spatial frequency. Several models of the CSF have been proposed in the literature (Mannos and Sakrison, 1974; Daly, 1992; Ahumada, 1996) and underpin many HVS models applied in areas such as error visibility assessment (Wang et al., 2004), image quality evaluation (Barten, 1999) or saliency (Perazzi et al., 2012). Mullen (1985) provides an extension of this work into colour vision by measuring human CSFs for the blue-yellow and red-green chromatic channels. She found that the absolute sensitivity and spatial resolution of chromatic channels is significantly lower than that of the achromatic channel. Furthermore, she notes the low-pass characteristic of the chromatic CSF in contrast to the bandpass shape of the achromatic CSF, which accounts for the higher absolute sensitivity of the chromatic versus achromatic channels for very low spatial frequencies. Further evidence by Losada and Mullen (1994) suggest that three

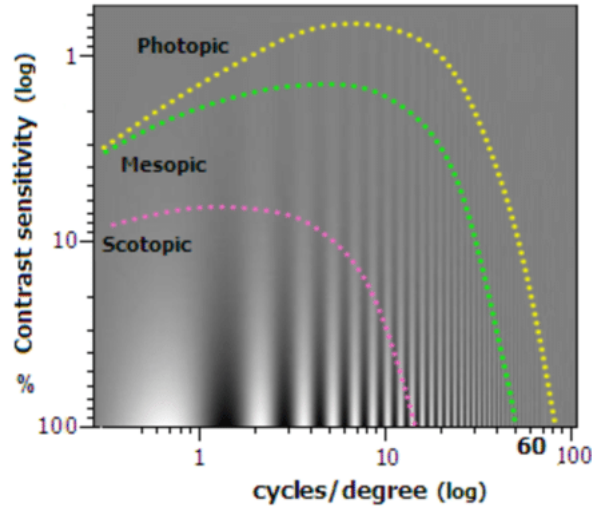


Figure 2.3: Empirical contrast sensitivity functions for scotopic, mesopic and photopic vision, measured for sine wave gratings at a single orientation overlaid on top of test stimulus - sine wave gratings at different spatial frequencies and contrast. Image courtesy of Arregui et al. (2020)



Figure 2.4: Illustration of the effect of decreasing relative contrast between background and target on the visibility of the target. The intensity of the text is progressively increased to match that of the background.

separate bandpass mechanisms tuned to different spatial frequencies are responsible for this. The spatial frequencies at which contrast sensitivity is measured are expressed in degrees of visual angle, implying that viewing distance has an impact on the visibility of certain details.

2.2.7 Visual Masking

Visual masking is a phenomenon that occurs when a stimulus, referred to as the *target*, is rendered perceptually invisible by the presence of another stimulus - the *mask*. This effect is rooted in the HVS's varying sensitivity to different spatial frequencies and orientations. Both spatial frequency and contrast have an impact on the degree of masking, e.g. for a given mask, the contrast at which the target becomes visible is referred to as the visibility threshold (Walter, Pattanaik and Greenberg, 2002). Legge and Foley (1980) investigated the impact of contrast and spatial frequency on the visibility threshold of one sinusoidal grating (the target) in the presence of another (the mask). They found that changes in both the spatial frequency and contrast of the mask had a non-linear impact on the visibility of the target. Specifically, low contrast masks had little effect on visibility thresholds. As the contrast of the mask was increased, the target visibility threshold initially decreased

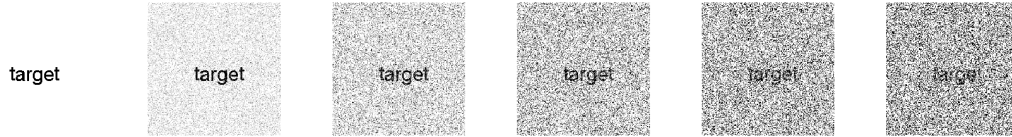


Figure 2.5: Illustration of the effect of visual masking. In this example, additive Gaussian noise of progressively higher amplitude is added to the original image (far left). In this scenario, the text is the target, whereas the noise is the mask. It is visible that an increase in the intensity of the mask results in a decrease in the legibility of the target.

at moderate contrast levels, before decreasing linearly for larger mask contrasts.

2.2.8 Just-Noticeable Difference/Distortion

The just-noticeable difference (JND), difference threshold, or difference limen, is a statistical property describing the amount by which a stimulus needs to be changed for a difference to be reliably noticed (Weber, 1996). More pragmatically, the JND is commonly defined as the amount of change to a stimulus that experimental subjects can detect at least 50% of the time.

This principle can be applied to measurement of a range of different physical stimuli, for example, particular types of image distortions. Due to properties of the HVS such as contrast and colour sensitivity, as well as effects such as masking, the HVS cannot sense distortions or appearance differences below certain thresholds, for a given spatial frequency. Distortion just above that threshold is referred to as the just-noticeable distortion and is dependent on both the local properties of the image, and the visual acuity of the observer. Distortions below the JND threshold can thus be disregarded when modelling perceptual image quality (Yang et al., 2005). JNDs are commonly measured subjectively using a signal detection paradigm. Additionally, some application driven JND models have been proposed, attempting to emulate subjective results for specific conditions and applications, such as image quality evaluation or video coding (Jia, Lin and Kassim, 2006).

Commonly, JND is measured in the spatial domain, and is affected by local properties of the visual stimulus, specifically *background luminance* and *masking* (Wu, Shi and Lin, 2019). Many JND models for visual stimuli explicitly correct for these effects (Chou and Li, 1995).

2.2.9 Visual Attention

Typical scenes encountered by the HVS contain many objects, constituting a large amount of visual information. Due to its limited capacity for information processing, the HVS does not treat all this information with the same importance. This often results in competition between objects in the visual field for neural representation. Consequently, the HVS exhibits a property of selectivity, which enables attended information to be processed and unattended information to be largely ignored, thus conserving the limited

processing capacity (Desimone and Duncan, 1995). Visual attention (VA) is the collection of mechanisms driving this selective behaviour and associated eye movements. The manner in which VA is allocated between the competing stimuli depends on both the intrinsic visual properties of the stimuli, known as *bottom-up* attention, as well as the task being performed by the HVS, known as *top-down* attention.

Bottom-up VA is dependent on the intrinsic features of visual stimuli, collectively referred to as their saliency (or saliency). Saliency describes the property of an object ‘standing out’ from its neighbours. Saliency and bottom-up VA are rooted in evolutionary theory and are linked to the facilitation of survival, by deploying VA to the most relevant information in a scene and responding to sudden threats, such as a predator, for example (Borji, Sihite and Itti, 2013).

In contrast, top-down VA is driven by a combination of the visual task at hand and, consequently, volitional deployment of attention. As this process is controlled by higher cognitive areas, it requires additional effort, compared to bottom-up VA (Itti and Koch, 2001). Top-down VA is employed in tasks such as visual search, where the goal is to find objects with some pre-specified properties. This allows for the HVS to attend to objects relevant to the task and disregard any irrelevant information. Both top-down and bottom-up VA mechanisms tend to operate in parallel in everyday scenarios, whereby attention is both modulated by the intrinsic properties of objects, and top-down properties such as memory, task and context.

Many approaches have been proposed to the task of computational VA modelling, particularly focusing on bottom-up attention and saliency, due to their relative simplicity and suitability for computational modelling based on the feature integration theory of attention (Treisman and Gelade, 1980). The goal of such models is to accurately predict highly-probable locations of human fixations in an image, thus pointing out attention-grabbing regions or objects in an image. Existing models tend to exploit either a local approach, using centre-surround differences computed in local regions, or a global one, assessing the entire image at once. While being biologically plausible, both approaches come with their distinct advantages and associated problems. For example, local methods tend to overestimate the saliency of object edges and high frequency image content, while missing larger salient regions, while global methods tend to work well for larger objects, but have trouble dealing with highly textured regions.

Notably, Itti, Koch and Niebur (1998) in their seminal work proposed a biologically-inspired bottom-up saliency-based model of visual attention. This model, given an input image, generates a corresponding saliency map through successive stages of colour, intensity and orientation feature extraction, centre-surround differencing and linear integration of these features at multiple scales. This model served as a foundation for multiple further developments, such as the graph-based saliency model of Harel, Koch and Perona (2007) or the conditional random field-based approach of feature combination

put forward by Liu et al. (2011). More recently, Wang et al. (2015) utilised deep neural networks to perform both local and global feature extraction and their integration into predictions of saliency.

Visual attention modelling plays a crucial role in machine perception. Similarly to the CSF, it allows for a perceptual weighting of image content to be computed.

2.2.10 Summary

This section has reviewed relevant background on human visual perception and discussed key approaches and conceptual frameworks used in vision research, highlighting the complexity of the HVS, as well as the benefits of taking a modular approach to its analysis. Additionally, a review of related work has been presented, reporting on empirical investigations into contrast sensitivity, visual masking, attention, scene analysis and distortion detection. The following sections build on these fundamental properties in the context of analysis and modelling of subjective visual properties of natural images, such as quality and realism.

2.3 Image Quality

Subjective properties of natural images, such as quality, realism, naturalness or beauty, are notoriously time-consuming and expensive to analyse and model. This is chiefly due to the sheer size and variance of the set of natural images, but also due to the inherent variability in human judgments, and the impact of experimental conditions during the collection of those judgments. These factors make it a non-trivial task to generalise models developed under experimental conditions to real-world scenarios.

The following sections discuss the concepts of *image quality* and *visual realism*, their definitions in different application areas, methods for measurement and modelling, as well as concrete examples of use in the domain of digital images.

2.3.1 Image Quality

Image quality is fundamentally related to the human experience of viewing an image. It can be thought of as an image characteristic measuring distortion or degradation, as compared to a reference image or a theoretical ideal image (an example is given in Figure 2.7). As humans have historically held the role of ultimate arbiters of image quality and are the end-users of most multimedia systems, many approaches to image quality assessment are developed and/or evaluated against human judgments as a baseline.

The process of image quality analysis by a human observer can be viewed as a special case of an observer interacting with their environment. This can be summarised as a cyclical process of (1) acquisition of environmental information and its internal representation; (2) cognitive interpretation and comparison of this internal representation

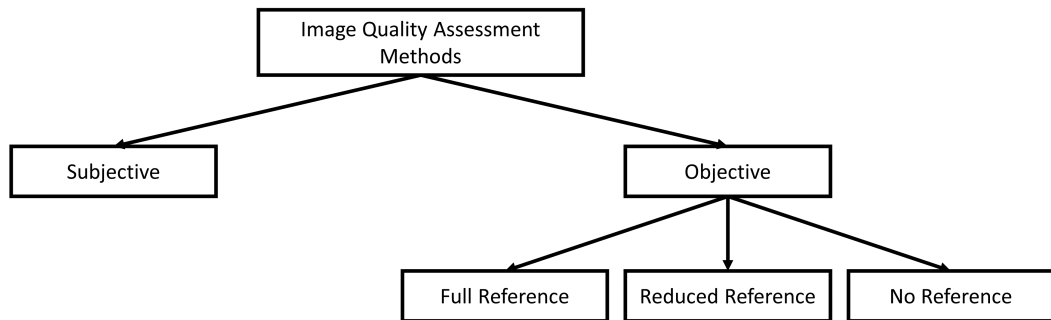


Figure 2.6: Overview of approaches to image quality assessment. Reproduced from Voronin et al. (2019)

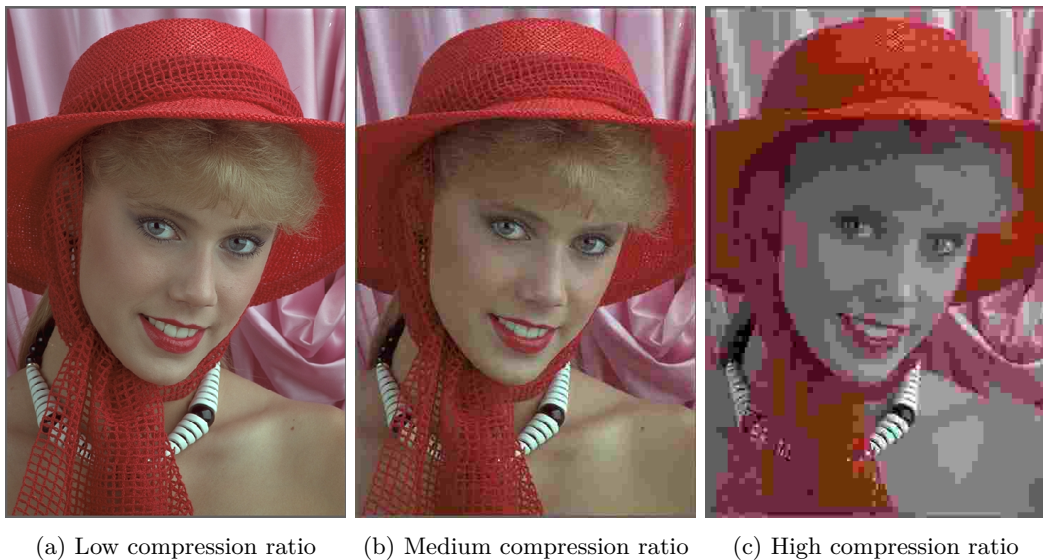


Figure 2.7: Images from the LIVE Database (Sheikh et al., 2005) illustrating the impact of JPEG compression on perceived quality.

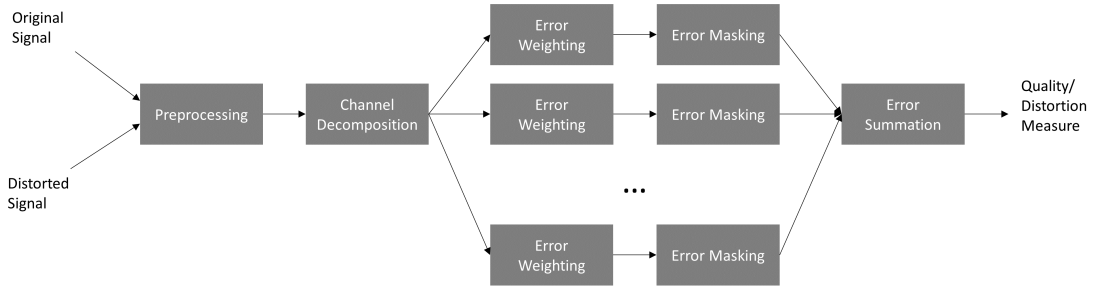


Figure 2.8: Illustration of the error sensitivity framework for image quality assessment. Image reproduced from Wang, Bovik and Lu (2002)

with relevant representations stored in memory and (3) an appropriate response, driven by this interpretation (Janssen and Blommaert, 1997). In the case of image quality, the image is the environment and its low-level features, such as resolution, contrast, brightness, colourfulness etc. are the acquired information. The internal representation is a higher-level interpretation of this information by an observer, and the response is the quantification of a subjective quality score based on this experience. Consequently, Leisti et al. (2009) argue that subjective quality ratings do not explicitly reflect the properties of the image, rather they reflect the subjective experience of quality based on these properties, as well as the individual differences between observers, such as their experience or expectations.

Alternatively, the overall quality of an image can also be seen as a function or combination of some set of features visible to and interpreted by a given observer (Kim et al., 2008). This is the foundation of error-sensitivity-based approaches to objective image quality assessment, illustrated conceptually in Figure 2.8. Such error visibility methods rely on modelling the perception of errors and their subsequent pooling into a quality score. These methods have been extended to structural-similarity-based approaches, which focus on the visibility of errors in areas of the image containing structural information. This was shown to produce quality scores which correlate more closely with human opinion scores, compared to error-visibility methods (Wang, Bovik and Lu, 2002; Wang et al., 2004). Alternative approaches also explored statistically-based frameworks (Sheikh, Bovik and Cormack, 2005), which relied on the assumption that the HVS has, throughout its evolution, become tuned to the statistical distributions present in the natural world. Accordingly, the disturbance of such natural image statistics produces a subject

Based on the above, it can be seen that definitions and measures of image quality fall into two key categories (Wang and Bovik, 2006):

1. **Subjective:** measuring aspects of the subjective viewing experience directly
2. **Objective:** assessing the image data itself using mathematical models to derive quality scores that correlate with human judgments.

These two approaches can be linked to the respective concepts of Quality of Experience (QoE) and Quality of Service (QoS), commonly used in information, communications and multimedia research. The former measures the end-user experience, expectations and perception, while the latter assesses the efficiency with which a service or system can deliver information without introducing distortion or loss (Fiedler, Hossfeld and Tran-Gia, 2010). Figure 2.6 shows a high-level classification of different approaches to image quality assessment.

2.3.2 Subjective Image Quality Assessment

Subjective image quality assessment (IQA) methods rely on collecting quality ratings directly from human observers, usually under strictly controlled experimental conditions. While many experimental procedures exist for this purpose, most share common principles:

- Observers are presented with a series of images for which they have to provide opinion scores
- Opinion scores are commonly allocated along a predefined scale
- Scores for different experimental conditions are averaged across subjects to generate *mean opinion scores* (MOS)

Subjective IQA procedures originate from experimental psychology, specifically from psychometric scaling methods (Torgerson, 1958) and have been standardised and categorised by the International Telecommunications Union recommendation BT-500, which sets out general experimental conditions, provides recommendations for experimental stimulus, observer selection, rating scales, results analysis and selection of test method (ITU, 2002). Key approaches to IQA and their variations and extensions are discussed further in Section 2.5.

2.3.3 Attributes Affecting Subjective Image Quality

Subjective perception of quality for a given image is heavily influenced by both its low-level (physical) and high-level (psychophysical) attributes. This is also affected by the sensitivity of observers to these attributes, their ability to map low-level features to high-level attributes, as well as the environmental conditions under which the image is viewed. Additionally, the context under which image quality is being evaluated, as well as the specific instructions given to observers, or their individual level of experience may further influence which image attributes are leveraged to arrive at final quality ratings.

Leisti et al. (2009) found that while different individual observers may vary with respect to the image attributes they rely on to rate quality, the reliability of quality ratings is high when those are averaged across observers. This suggests that while observers may adopt different subjective strategies, they tend to agree with respect to the final quality scores. The authors also identify a range of low- and high-level subjective attributes used

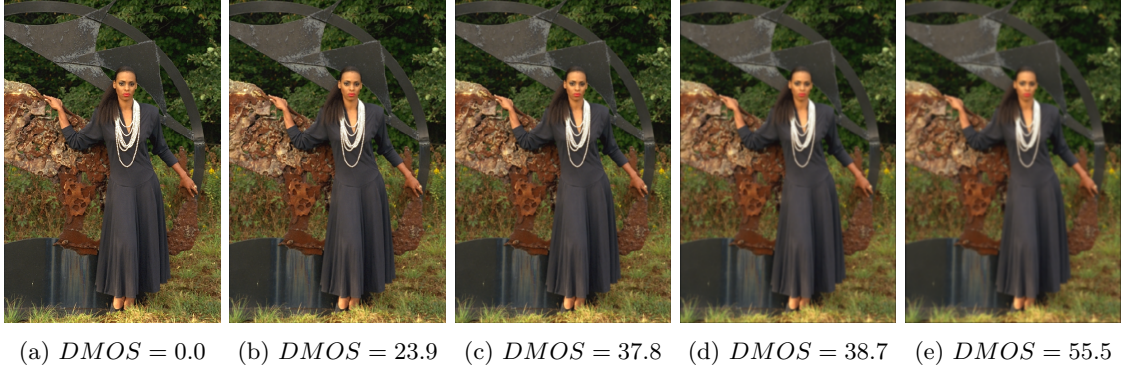


Figure 2.9: Example of the impact of Gaussian blur on the appearance of an image from the LIVE image quality database. The leftmost image has the lowest amount of blur, whereas the rightmost image has the largest amount of blur. Each subcaption shows difference mean opinion scores (DMOS), re-aligned from raw mean opinion scores. The DMOS can be seen increasing as the amount of blur is increased, indicating a decrease in subjective quality. See Sheikh, Sabir and Bovik (2006) for details.

by observers by means of post-test interviews. Amongst the most commonly used low-level attributes are *brightness*, *sharpness*, *lightness*, *brightness of colours*. *Realism/genuineness*, *naturalness*, *clarity* and *depth*, on the other hand, are the most common high-level attributes used by observers to interpret and translate low-level attributes to quality scores. This is further evidenced by Kim et al. (2008), who evaluate the impact of varying a range of low-level image attributes on subjective image quality ratings. Observers are tasked with rating images based on seven high-level psychophysical attributes: *image quality*, *naturalness*, *clearness*, *sharpness*, *contrast*, *colourfulness* and *preference*. The authors find significant correlations between multiple attributes, for example between clearness (sic), sharpness and contrast, or image quality, naturalness and preference. Ultimately, naturalness and clearness are found to be the two most significant attributes affecting image quality and are incorporated into a predictive model. In essence, it appears that these two high-level attributes provide a compact way to describe a host of low-level attributes relating to the spectral and spatial fidelity of an image (see blur example in Figure 2.9). Janssen and Blommaert (1997) also propose two key high-level attributes, but substitute *clarity* for *usefulness*, which describes how precise an image representation is. The authors define *naturalness* as the degree of correspondence between the representation of the image and the model of reality stored in one’s memory. This suggests that observers integrate low-level cues into higher-level semantic indicators of quality. It is important to note that the attributes related to subjective image quality may change depending on the stimulus and distortion type. Accordingly, this must be taken into consideration when evaluating experimental results.

2.3.4 Objective IQA

Objective image quality assessment methods avoid some drawbacks of the subjective approaches, by removing human subjects from the process and substituting them with

computational models of human visual perception, or more pragmatic metrics, which are approximations of some aspect of the human visual system. Depending on whether a reference image is used, objective quality metrics either calculate a distance between the test and reference images, or compare the features of the input image to an existing model or heuristic, depending on the specific aim of the quality evaluation. Approaches to objective image quality assessment can be broken into three main categories, according to presence of a reference image:

1. **Full-reference (FR)** methods, which assess image quality by comparison to a reference image assumed to have ideal quality;
2. **Reduced-reference (RR)** methods, which assess image quality by comparison to a set of features extracted from a reference image assumed to have ideal quality;
3. **No-reference (NR)** methods, which assess image quality without any explicit reference image

Objective IQA methods can also be categorised based on whether they measure quality in the context of **signal fidelity** or **human perception**. A review of objective IQA metrics is given in Section [2.6](#)

2.4 From Image Quality to Visual Realism

While image quality and visual realism relate to similar perceptual properties of images, they are fundamentally different concepts. Both can be used to describe the likeness of a visual representation to its original source, however, the contexts in which they are applied differ. Due to their origins in broadcast engineering, standard methods for IQA focus on measuring acquisition, processing, compression and transmission artefacts, often with reference to an original, unaffected image. The emphasis is thus on quantifying the accuracy of an optical and/or digital sampling process, as well as the degradation of a signal by a set of transmission or post-processing operations. In contrast, visual realism commonly describes the results of a synthesis or fusion of different elements, such as painting or 3-D modelling. It describes how well a synthetic approximation of reality compares to the experience or perception of that reality. This description can be performed at multiple levels of assessment and is often limited to the confines of the medium itself. For example, one can find a pencil sketch portrait realistic, despite the obvious limitations of lead on paper, such as the lack of colour information. At the same time, one can find a wax sculpture of the same subject unrealistic, despite the presence of additional features absent in the pencil sketch, such as true depth, scale, material etc. Interestingly, the relationship between the likeness of a representation to its target and the resulting subjective preference is not always monotonic. Specifically, subjective responses, or affinity, of humans towards realistic humanoid robots are subject to a sudden decline, when the robot appears very similar to a real human being. This effect is known as the uncanny

valley (Mori, MacDorman and Kageki, 2012). This suggests that techniques used for image quality assessment may not be well suited to predicting visual realism, or other subjective image properties.

This section describes the similarities and differences between image quality and realism, defines visual realism in the context of image compositing, and relates this to other approaches to image synthesis. Human perception of visual realism is also discussed, and the impact of different image features is reviewed. Existing approaches to subjective measurement of realism are then reviewed. This is followed by a discussion of related work in modelling and prediction of visual realism.

2.4.1 Limitations of Representation

Barbour and Meyer (1992) emphasise that the act of representing a dynamic three-dimensional environment, such as any natural scene, on a static two-dimensional piece of canvas, film, or paper poses several important issues, which make it impossible to achieve a perfectly realistic representation. Firstly, as opposed to a real scene, perceived directly with one’s eyes, a photograph, or painting is made from a single fixed point of view, as opposed to the two that our eyes offer. This removes binocular cues used by the HVS to extract depth information. Secondly, paintings and photographs are flat and have a fixed size and implicitly limited field of view, further limiting neural cues associated with depth perception. Moreover, neither paintings nor photographs are capable of conveying the same dynamic range as our visual system is capable of perceiving in the real world. Finally, all images are often seen under some kind of viewing illumination, which, through chromatic adaptation, has further impact on the final appearance. Consequently, a 2-D picture cannot convey all the same visual cues as a real-world scene, and thus truly physically-realistic pictures are theoretically unattainable. At the same time, this does not preclude the discussion of visual realism in the context of 2-D images, it simply constrains it to fewer relevant cues (see Figure 2.10). These constraints, due to the monocular nature of 2-D images, have forced artists throughout the centuries to come up with techniques to recreate the missing cues, given the limitations of the medium.

2.4.2 Definitions of Visual Realism

In recent years, more pragmatic studies of visual realism have been carried out, particularly in the areas of 3-D graphics and image compositing. The issue inevitably faced by the authors was the definition of realism in the context of these modern image generation tools. Ferwerda (2003) shows that evaluation of realism can be performed at three distinct levels of visual coding: physical, photographic and functional. *Physical realism* describes a circumstance where the image provides the observer with the same **visual stimulation** as the original scene, meaning that the array of spectral irradiance values incident on the retina during viewing of the original scene would have to be somehow reproduced in the presentation of the physically realistic image under evaluation. Currently, no such



(a) “Self-portrait” by Rudolf Wacker. Source: [Wikimedia / Public Domain](#)



(b) “Self Portrait #7” by Rob. Source: [Free-Images.com / Public Domain](#)



(c) “Portrait (pencil)” by Ricce. Source: [Wikimedia / Public Domain](#)

Figure 2.10: Pencil portraits of varying degrees of realism. Each of the portraits could be deemed as subjectively realistic, however, due to more detailed reproduction of certain cues, some could be deemed more realistic than others. While (a) is a plausible likeness of a human face, it reproduces certain cues with less detail than portrait (b). Arguably, portrait (c) is the most realistic, since it reproduces lighting, depth and shading cues with high fidelity and plausibility, compared to (a) and (b). Despite the common medium, each representation differs in terms of visual cues represented and the accuracy of their representation.

technology exists. In *Photo-realism* the image elicits the same **visual response** as the original scene. In the context of 3-D graphic or other synthetic content, one can call an image photo-realistic if it is indistinguishable from a photograph (e.g. Figure 2.11), or if it provides the same *photometric* information as the original scene, despite differences in the physical aspects of the stimuli (i.e. photograph vs real scene). Finally, *functional realism* describes an image that provides the same **visual information** as the original scene. This means that, while the style of depiction may not be faithful (for example using line sketches to depict human actions in an instruction manual for a piece of furniture), the information content is still conveyed (i.e. the actions taken by the sketched human beings in the instruction manual are still distinguishable). Cartoons rely particularly on this type of realism, thus being able to generate compelling storylines with mere visual approximations of the real world.

Since the definitions of visual realism and the properties they hinge on vary based on the method and style of depiction, it is useful to categorise them accordingly. Reinhard et al. (2013) proposes a categorisation into *four* distinct approaches to the process of synthesising visually realistic imagery: manual modelling, physical simulation, image-based rendering and data-driven synthesis. Manual modelling relies on the use of interactive 3-D modelling software in order to generate a realistic scene from scratch. This requires significant effort and does not guarantee realistic results, due to the degree of randomness present in the appearance of the real-world, compared to 3-D models. Methods based on physical simulation can generate very realistic results, for example simulating the physics of water or smoke, however are difficult to extend to other phenomena. Image-based rendering relies on capturing samples of the real world, creating a reconstruction of the captured world,



Figure 2.11: An example of a contemporary photorealistic painting: John Baeder, John’s Diner (2007) CC BY-SA (<https://creativecommons.org/licenses/by-sa/3.0>)

and then resampling that reconstruction based on the plenoptic function (Adelson, Bergen et al., 1991) to generate novel views of this environment. Finally, data-driven synthesis also uses samples of the real world, but with fewer constraints on how these can be modified and combined. Digital image compositing fits this category well, as it often reuses and modifies samples of the real world, by inserting novel content into them. Manual modelling and physical simulation are also of interest, due to their replication of physical processes that create the visual complexity of the natural world, such as simulating the interaction between lights and surfaces. Both these processes rely on a ground-up synthesis of objects and scenes and replication of global processes, such as illumination, reflection, and thus are linked to the domain of computer graphics. On the other hand, image-based and data-driven techniques reflect the photographic nature of the task of image compositing by using photographic samples of the real world and combining/manipulating them until a certain plausible effect is achieved. However, due to their inherent 2-D nature, they are limited to manipulation of 2-D image data, without direct access to the underlying 3-D structure. Specific data-driven synthesis techniques are further discussed in Chapter 4.

2.4.3 Perception of Visual Realism

In addition to the multiple methods of synthesis described above, what makes a consistent definition of visual realism challenging to arrive at is the multitude of visual features that create the perception of plausibility for the average human observer. These features can be coarsely broken into two categories: *physical* and *semantic*, following the approach of Biederman, Mezzanotte and Rabinowitz (1982). Physical cues relate to the plausibility of the physical relationships between objects in/and the scene, such as illumination, support

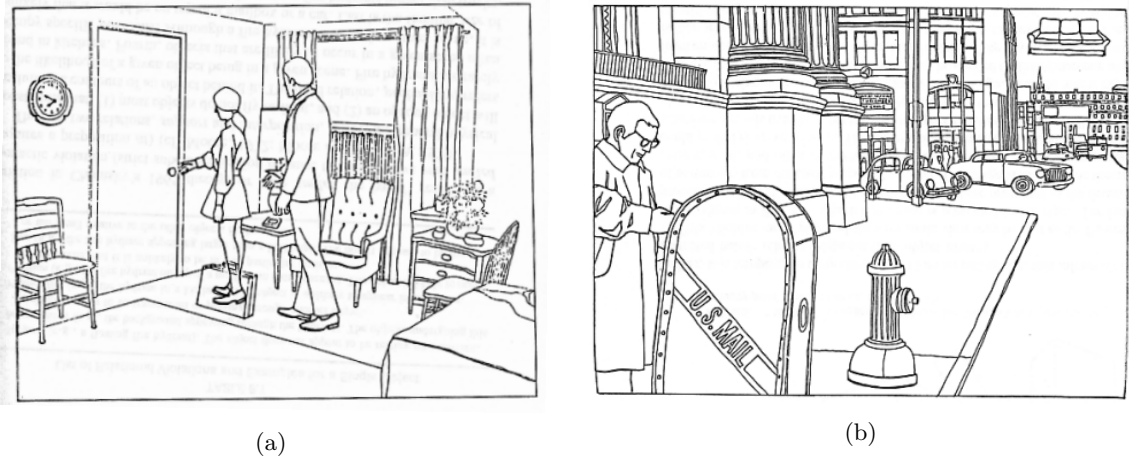


Figure 2.12: Examples of semantic and physical visual features undergoing *violations*. (a) A transparent briefcase: example of a physical cue violation (briefcases commonly occlude what is behind them), (b) “Goodyear Sofa”: Example of physical and semantic cue violation (sofas are usually smaller and don’t float over the street). Images courtesy of Biederman, Mezzanotte and Rabinowitz (1982).

or material reproduction. Semantic cues are related to the likelihood of seeing a given combination of objects, such as a horse riding another horse – even if the physical properties of this equine combination are rendered perfectly, the average observer would likely rate the realism low, due to the impossibility of seeing the aforementioned combination in the real world. Semantic cues are also likely to vary more from observer to observer, particularly if prior visual experience is a primary driver. Interestingly, some of these cues are very easy to spot for observers, while others can go unnoticed (Cavanagh, 2005).

It is well-known that some ecological short-cuts or assumptions are made by the HVS in order to efficiently interpret visual information, particularly the extraction of 3-D shape from 2-D retinal images (Ramachandran, 1988). For example, the ‘single light from above’ assumption, related to the extraction of shape from shading information (Kleffner and Ramachandran, 1992), suggests the HVS uses a weak prior assumption (Morgenstern, Murray and Harris, 2011) of a single light illuminating the scene from above (see Fig. 2.13a). The convexity bias (Liu and Todd, 2004), and generic viewpoint assumption (Freeman, 1994) are likely linked to the fact that humans tend to view objects from a specific subset of viewpoints and angles (see Fig. 2.13b). The existence of these assumptions suggests that not all visual information is processed in exactly the same manner by the visual system. The fact that some of these assumptions override others suggests that some hierarchy of importance must exist. Moreover, the disparity between the number of receptors in the retina (~130M) and the relatively smaller (~1M) number of axons in the optic nerve suggests the existence and necessity of processes which dramatically reduce information redundancy in early vision (Connors and Ng, 1989). Consequently, it is not surprising that the HVS can readily detect departures from reality in some of these features, such as unrealistic physical relationships between objects (Biederman, 1972), while completely disregarding others, such as physically impossible

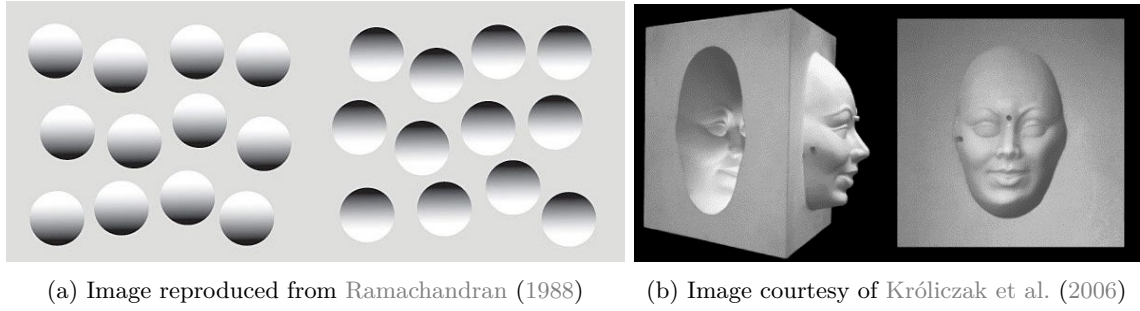


Figure 2.13: Two prior assumptions in vision illustrated. (a) “Single light from above”: Shading alone determines the percept of convexity in the circles on the left, and a percept of concavity in the circles on the right (b) “The hollow face illusion”: Even a concave face appears convex under certain viewing angles - this illustrates the effects of top-down processes in vision, enforcing a knowledge- or experience-based prior that faces tend to be convex. Interestingly, this prior seems to outweigh the assumption in (a).

reflections or shadows (Cavanagh, 2005; Reinhard et al., 2013).

2.4.4 Image Features Affecting Realism

There is evidence that a wide range of features have an impact on observer perception of visual realism. Biederman, Mezzanotte and Rabinowitz (1982); Biederman (1972) show that scene organisation and relationships between objects within a scene, specifically their support, interposition, probability of occurrence, position in scene and relative size have a significant impact on observer object detection and scene parsing and recognition ability. Interestingly, higher-order relationships between objects in a scene, such as illumination angle or shadow direction, do not elicit the same response and often go unnoticed (Ostrovsky, Cavanagh and Sinha, 2005; Cavanagh, 2005). Similarly, Vangorp, Laurijssen and Dutré (2007) showed how shape and material exhibit a reciprocal relationship: the shape of an object can change the perception of its material and vice-versa. Other image features may not directly affect perception of realism, but instead contribute to the detection of higher-level features, which, in turn may modulate the perceived realism of a scene. For example, Fleming, Dror and Adelson (2003) show that humans perform poorly at judging the semantic plausibility of the content of specular reflections (provided the illumination appears realistic). The authors suggest that the HVS uses assumptions about real-world illumination properties in order to estimate surface reflectance. Furthermore, reflections have a significant influence on our the correct identification of object shapes by humans (Fleming, Torralba and Adelson, 2004). In this scenario, *what* is reflected off an object is less important to the perception of a realistic object than the very presence of the reflection itself. Additionally, Pont and te Pas (2006) show that human observers tend to confound perceptual effects of illumination and material appearance in real-world scenes.

2.5 Methods for Subjective IQA

Before discussing existing approaches to measurement of visual realism, the following section discusses related work and approaches to measurement and modelling of subjective visual properties. Common approaches to experiment design, observer selection, stimuli presentation and viewing conditions are discussed and compared in the context of prior work.

2.5.1 Subjective IQA Measurement

Subjective quality evaluation methods rely on an experimental paradigm to collect subjective judgements from a group of human subjects under controlled experimental conditions. The overall aim of such methods is to evaluate image quality as perceived by a representative human observer, while minimising potential bias. Bias may arise due to:

- *experimental design* (e.g. order effects, such as fatigue or improving subject performance due to task practice)
- *experimental conditions* (e.g. display resolution, contrast, sharpness, ambient illumination etc.)
- *human subject variability* (e.g. low visual acuity, weak contrast vision, colour-blindness, experience, age, sex, mood etc.)

Subjective quality evaluation under controlled conditions provides representative results, however, tends to be time-consuming and expensive, due to the involvement of human subjects.

The key differences between existing subjective IQA methods are a) the manner in which stimuli are presented to observers during each trial of the experiment and b) the rating scale used for the collection of responses. Below, existing approaches to subjective image quality assessment and related work are discussed. Rating scales are briefly summarised, followed by a review of experimental procedures.

2.5.2 Stimulus Presentation

Aside from the adopted rating scale, the way stimuli are presented plays an important role in the structuring and categorisation of subjective IQA methods.

Single Stimulus

In single stimulus methods, a single *test image* is presented to observers, who are then requested to rate the quality of this image on a categorical, ordinal or continuous scale. No reference image is shown in the single stimulus method, thus it performs well in scenarios modelling real-world viewing conditions, such as video streaming, where end users' experience of quality is not based on direct comparison with a reference (Seshadrinathan

et al., 2010a). Due to their simplicity, single stimulus tests are relatively easy to implement and their results straightforward to analyse. Both the popular LIVE Image Quality Database (Wang et al., 2004) and LIVE Video Quality Database (Seshadrinathan et al., 2010b) adopted this approach, using a categorical and continuous scale respectively in order to capture baseline quality ratings.

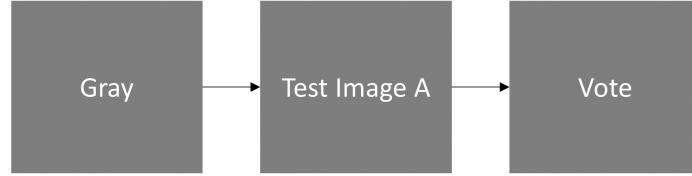


Figure 2.14: Single stimulus procedure, consisting of a grey screen, followed by the test image, followed by a voting phase.

Double stimulus

Double stimulus methods display two stimuli per trial: a *reference image* followed by a *test image*, although this order is sometimes randomised. Observers are either asked to evaluate the quality of each of the two images, or rate the quality of the test image, given the reference image. Sheikh, Sabir and Bovik (2006) used this approach to generate ground truth (GT) data for evaluation of full-reference objective IQA algorithms. Due to the requirement for rating of two images at a time (i.e. both the reference and test image), this method has been reported to yield more variable responses (Mantiuk, Tomaszewska and Mantiuk, 2012), compared to single-stimulus methods.

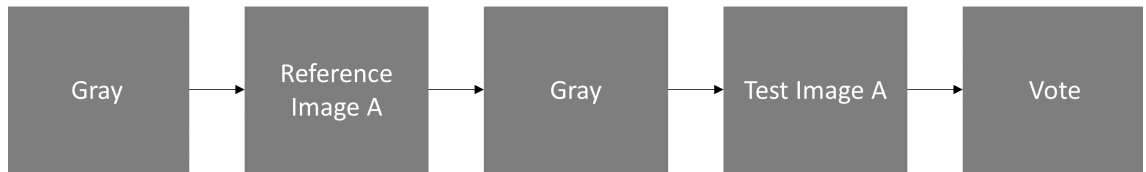


Figure 2.15: Double stimulus procedure, consisting of a grey screen, followed by the reference image, another grey screen, the test image, and finally the voting phase.

Forced-choice Pairwise Comparison / Simultaneous Double Stimulus

In contrast to the above methods, the forced-choice pairwise comparison approach presents observers with both the reference and test image at the same time, rather than sequentially. Under this paradigm, observers are required to select the image of higher quality, even if they cannot reliably detect a difference between the two images. The result of this approach is an ordering of test images according to their quality, as rated by observers. An extension of this approach is the **pairwise similarity judgments** method, which requires observers to both mark their preference, and indicate the magnitude of the difference between the two images presented in each trial. The forced-choice pairwise comparison (or

two-alternative forced-choice) test is also reported to be more time-efficient and produce most accurate results, compared to the single and double stimulus methods. Specifically, given $N = 2$ conditions, Mantiuk, Tomaszewska and Mantiuk (2012) found that the forced-choice comparison method required on average around 5s, compared to over 10s for the next best method (see Figure 2.17). In addition, using effect size comparisons, they showed that the forced-choice approach provided highest sensitivity and accuracy, compared to single- and double-stimulus approaches, as well as similarity judgements. Finally, this method was also shown to be easiest for observers to perform, requiring only a direct comparison of simultaneously visible images. These properties make it an efficient approach to adopt when large numbers of comparisons need to be carried out.



Figure 2.16: Double stimulus procedure, consisting of a gray screen, followed by the reference image, another gray screen, the test image, and finally the voting phase.

2.5.3 Grading Scales

Grading scales constrain the number and type of responses that observers can supply during experiments. Grading scales also differ depending on the measured property they pertain to. Ignoring unranked (i.e. nominal) qualitative descriptions, quantitative image quality grading scales can be coarsely split into two categories: **discrete** and **continuous**.

Discrete Scales

Perhaps the most commonly used discrete scales are the International Telecommunications Union (ITU, 2002) quality and impairment scales. These constrain the evaluation of quality and impairment perception to ordinal scales, such as the ITU-R BT (2002) five-point quality or impairment scale (see Table 2.1), or an n-alternative choice scale. Each point along the scale is assigned a semantically meaningful description, and the underlying number is not shown to subjects during the experiment.

While it is implied that the perceptual distances between consecutive points along these scales are a constant interval, or ratio (in the case of the impairment scale), Jones and McManus (1986) claim that this is not always true, particularly if the semantic labels are translated into other languages and recommend using numerical scales, instead of semantic ordinal ones, when the distance between ratings is important. Subsequent studies (Watson and Sasse, 1996, 1998) highlighted the relatively low reliability of this scale for problems where a high degree of accuracy is required.

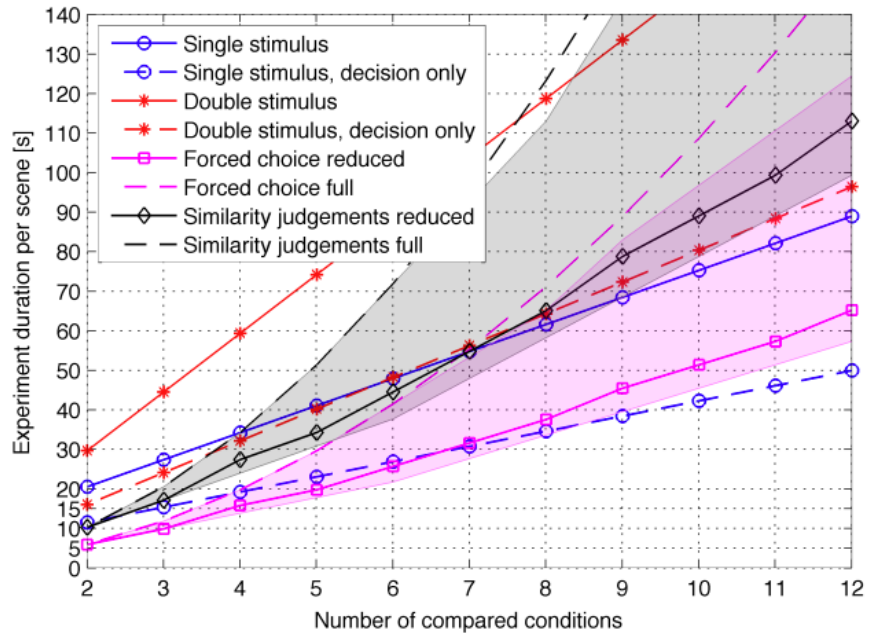


Figure 2.17: Average time required to compare N conditions under different experimental protocols. Image courtesy of Mantiuk, Tomaszewska and Mantiuk (2012).

ITU Five-grade scale		
Score	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 2.1: ITU five-grade quality and impairment rating scales. Reproduced from ITU-R BT (2002)

Continuous Scales

Continuous scales can be seen as providing observers with more flexibility in their responses, by not constraining them to a fixed set of pre-defined options. These scales commonly range between 0 and 1, or 0 and 100 and indicate key labels (usually the ones used in discrete scales) along the continuous scale. Common experimental methods which make use of continuous scales include the Single Stimulus Continuous Quality Evaluation (SSCQE), the Double Stimulus Continuous Evaluation Scale (DSCQS) and the Double Stimulus Comparison Scale (DSCS) (ITU, 2002). They are particularly useful for video sequences, where quality may vary along the temporal dimension.

In a four-way comparison study under the single stimulus paradigm, Huynh-Thu et al. (2011) found no significant differences between using discrete and continuous scales for video quality assessment, noting that observers tend to cluster their responses around the labels and tick marks of each scale used. This evidence is corroborated by Pinson and Wolf (2003), who suggest that observers rely on the same judgment process to map perceived errors to overall quality estimates under different continuous scale paradigms.

2.5.4 Experimental Conditions & Presentation

During an experiment, many factors can impact the appearance of an image to an observer and consequently, their responses. These include the display device, ambient and direct illumination, resolution, viewing angle and distance, to name a few. Similarly, the properties of observers, such as their visual acuity, experience or alertness can influence their own perceptions and responses. Finally, the manner in which images are displayed, presentation duration, size, location and order of the stimuli can all bias responses. Thus, when performing experiments, it is vital to constrain and normalise viewing conditions for all observers and conditions under test. The ITU Recommendation BT.500 (ITU, 2002) provides standard approaches for this, which stem from vision science and traditional psychophysics and have been adopted and extended widely in the field of IQA.

General Viewing Conditions

Environmental conditions under which image quality evaluation may take place can be coarsely organised into two scenarios: *laboratory* and *home*. The former is often used for accurate and highly controlled measurements, while the latter is supposed to account for the ‘wild’ conditions, under which end-users may view image content. In either scenario, the viewing conditions are summarised by calibration and measurement of the display device and the mathematical relationship of its properties with those of the room and the observer.

Display Device

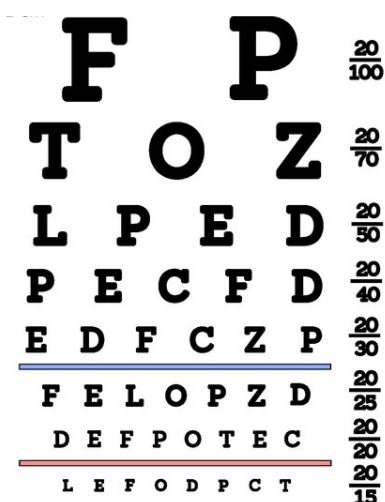
The display device used to render the images can also impact the viewing experience. Most commercial displays are capable of extensive colour calibration, which can significantly alter the appearance of viewed image content. It is thus imperative to perform calibration of the display device appropriate to the experimental setting. In order for experiments to be reproducible, standardised measures are used, providing references for the following properties of the display device:

- Calibration of brightness and contrast using standard greyscale test patterns
- Ratio of inactive screen luminance to peak luminance
- Ratio of peak luminance of screen displaying black, to that of screen displaying peak white in completely dark room
- Ratio of luminance of background behind screen to that of peak screen luminance
- Ambient illumination chromaticity
- Ambient illumination level
- Resolution
- Viewing Distance

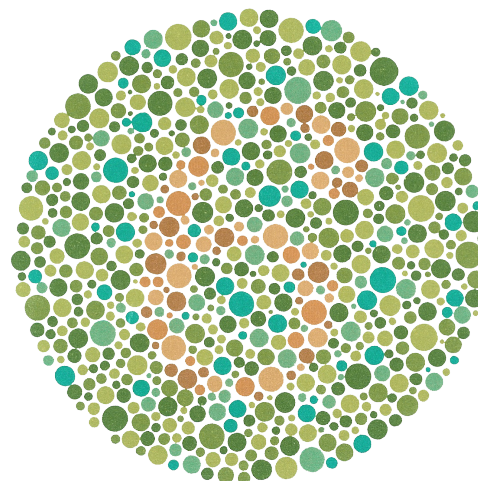
The device must also be capable of reproducing the full colour gamut of the experimental stimuli. Appropriate display device calibration and controlled viewing conditions ensure reliability of the results and reduce response variability. However, as mentioned earlier, this is not representative of home viewing environments, where larger response variability can be expected, due to the lack of control over the above variables.

Test Materials

The content of image-based experimental stimuli has been shown to affect subjective quality judgments. Jumisko, Ilvonen and Vaananen-Vainio-Mattila (2005) have shown that a connection between subjective interest in particular content and associated quality scores exists. Similar results have been obtained in experiments with audiovisual content, where modality (auditory, visual, audio-visual), interest and dynamics of camera/content had significant impact on subjective mean opinion scores (Lassalle et al., 2012). The order and type of stimuli, as well as the scale used to grade them, have also been shown to bias responses (Corriveau et al., 1999). In order to minimise such *contextual effects*, the selection of test materials (i.e. images) and the distortions or transformations applied to them (e.g. compression) must be carefully considered and appropriate for the selected experimental procedure and variables under investigation. Additionally, the content of images used in the experiments should be reflective of the task at hand. This is often achieved by either sampling a large enough number of images to diffuse the effects of individual images, or constraining evaluation to images of particular content (i.e. images of particular scenes or objects). Presentation order randomisation, as well as repeated



(a) A Snellen chart. Source: [Wikipedia, Public Domain](#)



(b) Ishihara colour blindness test, Plate No. 13 with an orange number 6 on a green background. Source: [Wikipedia, Public Domain](#)

Figure 2.18: Examples of standard test charts for (a) visual acuity and (b) normal colour vision.

stimulus presentation, are also effective approaches to minimising contextual effects (Choi, Jung and Jeon, 2009).

Observers

Appropriate selection of observers is paramount to the validity of any image quality assessment task. Due to the large number of possible variations between observers, such as their visual acuity, experience in particular image or distortion types, their motivation, age, gender, cultural background and occupation can all have an impact on results (ITU, 2002). In practice, it is very difficult to control for all these variables simultaneously. Nevertheless, care should be taken in order to balance observers both within and across groups, based on these properties. The ITU recommend using at least 15 observers for formal studies, however, point out that this also depends on the sensitivity and reliability of the test procedure.

Furthermore, all observers must be screened for normal visual acuity (Holladay, 2004) and colour vision (Birch, 1997). Standardised tests exist for both visual acuity (e.g. the Snellen chart Snellen (1868) or the “tumbling E” test (Keeffe et al., 1996)), and colour blindness (Ishihara et al., 1918). Examples of such standardised tests can be found in Figure 2.18. Observers must also be provided with clear instructions and be naive to the purpose of the tests, unless required otherwise. Opportunities to ask questions should be provided in order to normalise responses across the observer sample.

Test Session

In order to minimise bias due to observer fatigue, the tests should be limited to around 30 minutes at a time (ITU-R BT, 2002). For significantly longer designs, sessions should be split into multiple shorter parts. It has been shown that short training sequences, as well as stabilising sequences, results for which are not considered in the evaluation, are effective at accommodating observers with the specific task and apparatus, thus preventing order effects (ITU-R BT, 2002). In scenarios where control of the experimental conditions is difficult, such as remote users, detailed screening and evaluation of observers should be performed and incentives can be provided to increase engagement and task completion (Ribeiro, Florencio and Nascimento, 2011).

2.5.5 Analysis of Results

Due to the large variance of responses of individual observers to a particular stimulus, as well as the variance between observers, it is common to aggregate scores across observer groups, as well as multiple viewings of the same stimulus for individual observers. The *mean opinion score* (MOS) is a generalised concept referring to the arithmetic mean of all opinion scores for a given stimulus provided using some pre-defined scale

$$MOS = \frac{\sum_{n=1}^N R_n}{N} \quad (2.1)$$

where N is the number of subjects and R refers to an individual opinion score for a given stimulus.

The MOS has been standardised by the ITU and successfully applied in subjective quality analysis in different domains, such as speech, audio, images and video (Streijl, Winkler and Hands, 2016). Due to its conceptual simplicity, provided ITU recommendations for the minimum number of observers are followed, the MOS can be easily adapted to different scales and stimuli.

2.5.6 Other Comparative Frameworks

Subjective methods based on the above techniques have also been applied to evaluation of other image-based properties. For example, (Banterle et al., 2009) design a 2AFC-based experiment to facilitate subjective evaluation of different inverse tone-mapping operators. Similarly, subjective techniques have been used to compare the quality of HDR video compression (Mukherjee et al., 2016) algorithms, or to optimise resource allocation in multi-modal virtual environments (Doukakis et al., 2019). While these frameworks do not deal directly with measurement of the same subjective properties, many aspects of experimental methodology are very much relevant to this work.

2.6 Methods for Objective IQA

As introduced in Section 2.3.4, objective IQA metrics can be used to mitigate some issues associated with subjective methods. This section outlines key approaches to development of objective IQA metrics in the context of related work.

2.6.1 Signal Fidelity Measures

IQA methods based on signal fidelity assess the impact of various processes such as transmission, broadcast, storage, display, compression on the quality of images or video by means of statistical comparisons of the input and output signals. One of the most commonly used metrics for assessing signal fidelity is the mean squared error (MSE) calculated on the intensities of the input and output image pixels.

$$MSE(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2 \quad (2.2)$$

Here, x and y are the two images to be compared. This metric has been used extensively in the comparison of signal processing systems and their impact on signal quality. This is largely due to its attractive properties, such as its mathematical simplicity and ease of implementation. Moreover, its properties of non-negativity, symmetry, convexity and differentiability make it a very useful cost function for optimisation tasks (discussed further in Section 3.2). Furthermore, MSE allows for direct comparisons of similarity and interpretation in Euclidean distance space.

In scenarios where images of different dynamic ranges are compared, MSE is commonly converted to peak signal-to-noise ratio (PSNR).

$$PSNR = 10 \log_{10} \frac{L^2}{MSE} \quad (2.3)$$

Here L is the dynamic range, expressed as the number of possible pixel intensities, e.g. for an 8 bit image $L = 2^8 - 1 = 255$.

Despite its many advantages, it is well known that MSE is a poor predictor of human quality judgments, as it does not take the properties of the HVS into consideration (Teo and Heeger, 1994; Eskicioglu and Fisher, 1995; Wang and Bovik, 2009). (Lin and Kuo, 2011) claim that this is because not every difference between the reference and distorted image is noticeable by human observers, their attention is not deployed equally across the image. Similarly, some changes to image content do not correlate with a decrease in quality. For example, many post-processing operations which directly change pixel intensities, such as denoising or edge sharpening, aim to improve, rather than degrade, the quality of an image. Finally, due to spatial, temporal or chrominance masking, the

magnitude of intensity differences between the reference and distorted image is not always commensurate with the magnitude of change in quality, as perceived by a human observer.

2.6.2 Perceptual Visual Quality Metrics (PVQM)

In order to address the issues associated with signal fidelity-based measures, several methods which leverage known properties of the HVS have been developed. PVQMs can be broadly split into two categories: those which attempt to systematically model relevant properties of the HVS (vision- or model-based); and those which measure how pronounced certain features related with image quality degradation are, and predict quality scores based on this information (signal-based).

Vision-based PVQMs

Vision-based PVQMs rely on heuristic modelling of the known properties of the HVS in order to weight distortions in images by their visibility to human observers. Specifically, these models involve decomposition into spatial, colour and/or temporal channels, the contrast sensitivity function (CSF), luminance adaptation and a range of masking effects. Various approaches to modelling the HVS in IQA have been proposed, and many share key fundamentals. The visible differences predictor (VDP) proposed by Daly (1992) is one such approach. This FR method consists of three stages: calibration, HVS modelling and difference visualisation. The calibration stage parametrises the viewing distance from which the model is to make a prediction as well as properties of the display device, such as pixel spacing. This is followed by the HVS model, which models variations in visual sensitivity as a function of luminance, contrast and signal content. The output of this stage is a detection probability map for the reference and distorted images. The final stage consists of generating and visualising the difference between these two probability maps. This approach has been since extended to high dynamic range images (Mantiuk et al., 2005, 2011), incorporating the modelling of light scattering in the optics, local adaptation and non-linearities in luminance response.

Many existing vision-based PVQMs follow a similar high-level structure to Daly’s work, but adopt different rules for colour space, spatio-temporal decomposition, error pooling and visualisation. For example, (Lubin and Fibush, 1997) incorporate a Gaussian pyramid when performing decomposition, in order to simulate different spatial frequency bands. While these models have improved over approaches such as MSE, their complex design makes them expensive to compute and their reliance on often incomplete knowledge of the HVS makes them difficult to generalise to real world stimuli.

Signal-based PVQMs

Signal-based PVQMs attempt to avoid the issues associated with Vision-based PVQMs by replacing HVS models with the extraction and analysis of perceptually-relevant signal fidelity criteria, such as visual information or specific artefacts. They build on FR

approaches such as MSE and PSNR by comparing higher-level properties of the images, which are known to correlate with visual perception and consequently image quality. Under this paradigm, Sheikh, Bovik and De Veciana (2005) propose an information-theoretical approach using natural scene statistics, which models image distortions as information bottlenecks and expresses signal fidelity, as the amount of information a test image contains about a reference one. This work is then extended by Sheikh and Bovik (2004) through implementing a visual information fidelity measure (VIF). VIF quantifies information contained in the reference image and combines this with the amount of this information that can be recovered from the distorted image. Wang et al. (2004) exploit the importance of structural information in human visual scene perception and propose the Structural Similarity Index Measure (SSIM), which rates image quality based on the degradation of structural information between the reference and distorted images. SSIM is a weighted combination of three measures: luminance, contrast and structure. Many NR signal-based PVQM measures focus on specific types of distortions such as blur, sharpness or compression artefacts. For example, Marziliano et al. (2002) approximate perceptual blurriness by measuring widths of vertical edges in the image, while Caviedes and Gurbuz (2002) use local edge kurtosis as a proxy for perceptual sharpness.

2.7 Methods for Measuring Realism

Visual realism impacts systems well beyond aesthetic aspects. There is evidence that human task performance can be affected by the visual realism of a virtual environment or other task-specific image stimuli. This effect has been studied, particularly in the context of visual search (Lee et al., 2013; Ragan et al., 2015) as well as navigation in virtual spaces (Meijer, Geudeke and Van den Broek, 2009; Lokka et al., 2018). The performance increase noted in more realistic VEs is sometimes explained by the subjective increase in presence, which is often linked to task performance in VEs (Welch et al., 1996). It is important to mention, however, that higher visual realism does not always correlate with better task performance. Smallman and John (2005) argue that in many cases naive reliance on highly realistic visual displays can be detrimental and provides evidence based on geospatial data interpretation. This is corroborated in a subsequent study of visually realistic map renderings, which result in longer navigation task completion times and lower task accuracy compared to abstracted, less realistic line drawings of the same data (Wilkening and Fabrikant, 2011). Since visual realism is not a universal concept, and its impact on task performance is heavily modulated by the task and related stimuli themselves, the need for efficient approaches to its measurement and modelling is clear.

To date, there have been few attempts to quantify and model visual realism. Most current approaches evaluate the impact of a single, or a handful of features on subjective realism within a constrained image set. For example, Rademacher et al. (2001) varied shadow softness, surface smoothness, number of objects in scene, variety of object shapes and number of illuminants in both photographs and computer-generated scenes and measured

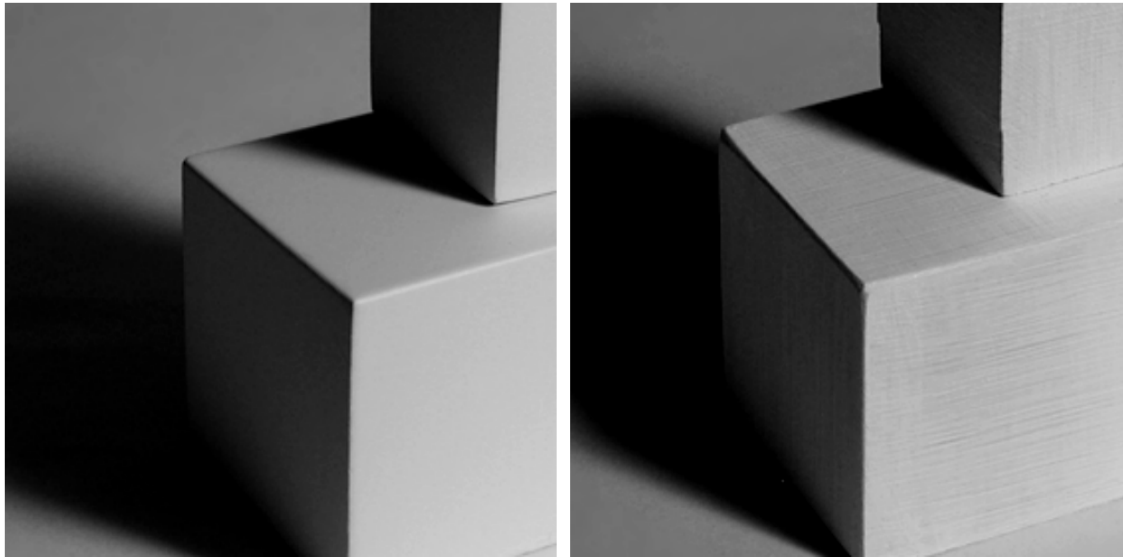


Figure 2.19: Experimental stimuli from Rademacher et al. (2001) varying surface smoothness from smooth (left) to rough (right). Experimental subjects found the shapes with rough surfaces significantly more realistic, compared to the smooth-surface shapes

the impact of each on subjective realism perception. The authors chose to simplify the scene, relying on simple 3-D volumes, plain backgrounds and no colour information. The test procedure adopted was a single stimulus discrete scale procedure, meaning each observer was presented with a single image at a time and was asked to rate it as either 'real' or 'synthetic'. Surface roughness was found to have the largest positive impact on realism of all parameters under test, followed by shadow softness. Interestingly, neither number of objects, variety of object shapes nor the number of lights had a significant impact on subjective realism, leading to the conclusion that in practical scenarios, efforts are better focused on shadow softness and texture, over scene arrangement and additional illuminants. This work was later adapted by Wang and Doube (2011) who employed these findings as hand-crafted features in their perceptually-inspired realism predictor for computer game images. Specifically, their predictor relied on measures of surface roughness (gradient variance), colour variance and shadow softness. This approach achieved moderate success in ranking the realism of real images higher than computer games, however the use of simplifying assumptions prevents this approach from being robust to outliers.

In contrast, McNamara et al. (2000, 2005) compared subjective realism responses to a physical 3-D scene, photographs, as well as 3-D renderings of the same scene. The authors opted for a lightness matching paradigm, which relies on observers utilising 3-D geometry and lighting information to make their judgments. It follows that a scene with inaccurate geometry and illumination would lead to less accurate lightness estimates by observers, while the real physical scene would yield most accurate estimates. While this approach only provides relative estimates of visual realism, it proposes a robust framework for evaluation of visual realism based on well-established visual cues and a task-based

experimental design.



(a) Low realism images



(b) High realism images

Figure 2.20: Example of the output of the visual realism model by Lalonde and Efros (2007). Red borders indicate unrealistic composite images, blue indicate realistic composite images and green indicate real photographs. (a) images deemed least realistic by the model, (b) images deemed most realistic by the model.

In the domain of image compositing, image-statistical approaches have found significant adoption, particularly due to the reliability of statistical properties of natural images. Torralba and Oliva (2003) showed how different object, scene and object+scene image categories can be distinguished based on second-order image statistics. This property was later exploited in scene recognition, specifically to build the *gist* of a scene by leveraging its global image statistics (Oliva, 2005; Oliva and Torralba, 2006). In the domain of visual realism assessment, Lalonde and Efros (2007) analysed colour distributions in natural images and proposed two approaches to evaluating their realism using colour compatibility. The first approach relies on comparing the colour statistics of an image to global colour statistics derived from a large dataset of natural images. This can be achieved by modelling natural colours of photographs as a single distribution, finding co-occurring background colours, given some foreground colours, or finding nearest neighbour scenes, given an object and adopting their colour distribution statistics. A local method is also proposed, which compares the colour distribution of the object and its background. A combination of both methods was eventually found to perform best. This approach was extended by Wong et al. (2012) who exploited local differences between colour distributions of unrealistic and realistic composite images. Using local colour statistics sampled around the interface between a background and a composited object, they calculate a colour similarity measure based on the intersection of histograms of the foreground and background regions. They combine this metric with *colour tendency* - a measure based on two properties: colour linearity and greyness, which traces the dominant hue in an image region, based on the heuristic that unrealistic composites tend to have diverging colour tendencies. These metrics were then implemented as features in a visual realism classifier and a recolouring algorithm.

A different image-statistical approach to automatically adjust features of image composites and improve their realism was proposed by Xue et al. (2012). The authors identified multiple key statistical properties, highly correlated between the foreground objects and background scenes of image composites: luminance, correlated colour temperature, saturation and local contrast. The impact of mismatches in these properties on subjective realism ratings was also evaluated. Finally, for a given combination of foreground object and background scene, the proposed algorithm minimises the object-scene mismatches for each feature by shifting and aligning histograms. Interestingly, in order to ensure that only one variable is changed, the authors used real photographs, which have been segmented into an object and scene part and subsequently processed in a manner such that a feature mismatch is introduced between the object and scene.

An alternative approach was adopted by Fan et al. (2014), who presented a large crowdsourced study of visual realism in real and computer-generated imagery. Two key contributions were made here: 1) an extensive, human-annotated dataset for the study of visual realism; and 2) results of psychophysical analysis of higher-level image attributes correlating with visual realism, such as the property of appearing to be a photograph,

natural lighting or observer familiarity, to name a few. The authors also proposed a computational model for the estimation of visual realism, based on their experimental results. This work was extended by Fan et al. (2018) who evaluated a range of state-of-the-art classifiers and found that modern ones, such as Support Vector Machines, Multilayer Perceptrons and Convolutional Neural Networks, when coupled with empirical features, all achieved similar performance at binary realism classification of images from their dataset. Hand-crafted features such as GIST (Oliva and Torralba, 2006), SIFT (Lowe, 1999), HOG (Dalal and Triggs, 2005) and LBP (Ojala, Pietikainen and Maenpaa, 2002) fed into a support vector machine all scored significantly lower, further underscoring the inherent value of empirical features in generalising visual realism. Despite this success, human performance was not surpassed.

Some researchers have also explored the question of realism from a signal-, or error detection perspective. For example, studies by Ostrovsky, Cavanagh and Sinha (2005); Koenderink, van Doorn and Pont (2004); Lopez-Moreno et al. (2010) evaluated human sensitivity to illumination direction inconsistencies and found that observers could not reliably detect inconsistencies of less than $20 - 30^\circ$. These studies all used a similar approach, whereby multiple versions of the same shape were shown in an n -alternative forced choice task, one of which was illuminated from a different angle compared to the $N - 1$ alternatives. Observers were tasked with finding the oddly illuminated object from amongst the distractors. Besides the angle of illumination, these studies also show further impact of texture, shape complexity and number of objects on observers' ability to accurately detect oddly illuminated objects, highlighting the large tolerance for inconsistencies and a serial visual search process underlying this behaviour.

2.7.1 Realism & Visual Attention

Attempts to understand the process of subjective realism assessment and the relative importance of image features have been also made by Elhelw et al. (2008), who used an eye-tracking paradigm to analyse how visual attention was deployed by observers assessing realism of real and computer-generated bronchoscopy images. In order to achieve this, an initial gaze-based study revealed multiple salient features through a single-stimulus realism rating experiment. These salient features were then modified in a second experiment,



Figure 2.21: Example of statistical composite harmonisation. Original source and target image shown in leftmost panel. The second panel illustrates a combination of the object from the source image and the target scene. The subsequent panels show results of harmonisation using three different algorithms. Image courtesy of (Xue et al., 2012)

where a realistic CGI version of the original images was created, allowing for the properties of the salient features to be modified. In turn, this enabled for each of the salient features to be ranked with respect to its importance in generating a visually realistic perception in observers. The authors found that specular reflections and geometric cues, such as silhouettes and shadows around edges, were the two most important features. Moreover, they illustrated the importance of visual attention in the subjective realism assessment process. However, Ninassi et al. (2007) point out that despite the importance of visual attention, its modelling in the process of image quality assessment is a challenging task. Their attempts to improve objective IQA methods using attention models did not result in significant improvements, even when using empirical ground truth data. The role of visual attention in subjective quality and realism assessment is further discussed in the context of Chapter 5.

2.8 Summary

This chapter has presented an overview of key concepts and literature pertaining to measurement of subjective properties of images, particularly image quality and visual realism, as well as discussing relevant background on human visual system, its modelling and incorporation into image metrics. The concept of visual realism has also been defined and posited as a special case of image quality. Based on a literature review, the concept of photographic realism (or photorealism) has been reviewed and discussed in the context of image compositing. A review of related work in realism perception, measurement and modelling has also been presented, showing that many image-statistical approaches perform well as proxies for realism prediction, but are difficult to generalise and extend to new stimuli. Finally, related work discussing the impact of different visual features on the perception of realism has been presented. The importance of modelling observer attention has also been highlighted, showing that observer interest, experience, as well as the task at hand may influence the realism rating process, and thus should be incorporated into subjective models of visual realism.

Importantly, this chapter has illustrated that objective proxy metrics for subjective properties such as quality or realism, while more efficient and easier to adopt in practical scenarios, often under-perform in predicting human visual performance, compared to subjective methods. Subjective models, on the other hand, are commonly impractical to develop, particularly at scale, due to the requirement for controlled experimental conditions and human observers. This indicates that techniques for conditional generalisation of empirical perceptual models to novel image content could combine the perceptual relevance and task specificity of subjective approaches with the ease of use and efficiency of objective metrics. If the subjective assessment process is viewed as a function mapping from input images to subjective opinions under a certain set of task constraints, recent developments in learning-based methods could be leveraged to approximate the function performed by human observers. Accordingly, the following chapter presents a

review of recent advances in machine learning in this context.

Chapter 3

Advances in Machine Learning for Image Quality

3.1 Introduction

Chapter 2 illustrated that measurement of subjective perceptual properties of images such as quality or visual realism is a complex problem. Despite the existence of a range of approaches to image quality assessment, many are limited in terms of generalisability to different stimuli, degradation types or observer groups. The viability of objective metrics based on handcrafted features is largely reduced by the complexity of the HVS, as well as the inherent variability of subjective judgements and natural visual stimuli. Subjective measurement, on the other hand, is time-consuming and challenging to perform at scale, requiring lengthy experimental sessions, large observer groups and extensive result analysis. In recent years, developments in the field of machine learning have found application in many areas of pattern recognition and image analysis. These approaches offer a route to combining the accuracy of subjective approaches with the efficiency and scalability of objective metrics.

This chapter provides relevant background on machine learning and presents a literature review of recent advances in applying learning-based models to approximating perceptual functions, modelling observer performance and learning feature distributions useful for image quality and realism assessment.

3.2 Background on Machine Learning

3.2.1 Introduction

Machine learning (ML) is an interdisciplinary field of computer science, with strong links to statistics, optimisation, game theory and information theory. The aim of ML is to design computer programs that learn from experience, instead of relying on manually specified

instructions. This is particularly useful for complex tasks, difficult to explicitly describe in a rule-based manner, for example image classification or speech recognition. While such tasks are intuitive and straightforward for humans to complete, they have posed a significant problem for computers for many years. Rules for such systems are extremely difficult to specify formally, as they rely on complex, hierarchical combinations of simpler features and are heavily influenced by local context. For example, the intensities of pixels in an image of a cat can take on a near-infinite number of configurations, while still appearing like a cat to a human observer. Occlusion, illumination, or camera angle are just a few properties, which can completely change the numerical representation of an image, without affecting the semantic content at all. Regardless of viewing angle or illumination, a cat is still a cat. As such, a formal definition of all pixel intensity configurations mapping to a semantic concept of a cat is inefficient, if not impossible, to design by hand. Consequently, the concept of solving such problems by learning from experience – has long motivated ML researchers, particularly given the exponential increase of available data as a result of the proliferation of the internet.

While it is beyond the scope of this work to provide an exhaustive historical overview of machine learning, a key distinction between classical and contemporary approaches must be made. Classical machine learning algorithms, such as logistic regression (Cox, 1958), linear regression (Stanton, 2001) or k-means clustering (Hans-Hermann, 2008), rely on hand-designed features. If appropriate task-relevant features are designed and extracted from training data, these algorithms perform very well, in addition to being computationally inexpensive. However, when applied to complex tasks, such as speech recognition or image classification, manually engineered features have proven ineffective. Contemporary machine learning techniques, such as deep artificial neural networks (DNNs), overcome this limitation by incorporating feature extraction into the learning problem. This wider concept of first learning a task-relevant feature representation and then learning the mapping from that representation to the desired output is referred to as *representation learning*. At the time of writing, *deep learning* methods are achieving state-of-the-art results in a range of machine learning problems, through hierarchical representation learning, which expresses complex features as a combination of simpler features. This section summarises the key concepts behind machine learning, focusing on deep learning in particular, due to its successful applications in a range of pattern recognition problems.

3.2.2 The General Learning Problem

Formally, ML algorithms focus on solving what is defined by Mitchell et al. (1997) as the general learning problem:

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .”

The task T describes the function to be performed by the learning algorithm. It specifies how a collection of **features**, representing a **training example**, is processed. Features are numerically expressed properties of an object or event, e.g. the intensities of pixels in an image, daily market prices of a stock, or physiological measurements. A training example is thus expressed as a vector $\mathbf{x} \in \mathbb{R}^n$, where each element x_i represents a feature. A collection of training examples is referred to as a dataset $\mathbb{X} \in \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$. In supervised learning problems, a dataset consisting of target values $\mathbb{Y} \in \{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)}\}$ is also specified. Each example $\mathbf{x}^{(i)}$ will have a corresponding ground truth value $\mathbf{y}^{(i)}$. The dataset thus constitutes the experience E , required for the ML model to improve performance on task T . However, in order to quantify the performance, a performance measure P must be used. This measure is commonly specific to the task and measures the degree of departure from the correct answer or desired outcome.

Various tasks can be addressed with ML algorithms, a few common examples are presented by Bengio, Goodfellow and Courville (2017):

- classification
- classification with missing inputs
- regression
- transcription
- machine translation
- structured output
- anomaly detection
- synthesis and sampling
- imputation of missing values
- denoising
- density / probability mass function estimation

A simple example of a classification task applied to image data could be identifying whether a given photograph contains a cat. Such a task is defined by the user and must be appropriately reflected by the training experience E . An example could be a dataset of images representing input features, and associated binary ground truth labels describing feline presence in each photograph. In order to measure the accuracy of this classifier, the proportion of correctly classified photographs could be used as a performance metric P . Thus, whichever ML algorithm one chooses to apply to this problem can only be considered effective if the performance metric P improves as a function of experience E . In the context of the cat classifier, this means that as more examples are presented, the proportion of correctly classified images increases.

3.2.3 Maximum Likelihood Estimation

Many ML techniques rely on various parametric families of probability distributions. In order for ML models to perform well in a given task, the parameter values $\boldsymbol{\theta}$ of these distributions must be appropriately set. A common principle used for this purpose is maximum likelihood estimation (MLE). MLE finds the set of parameters which maximise the likelihood that the observed data was produced by the process described by the model. Accordingly, following the formulation of Bengio, Goodfellow and Courville (2017), the maximum likelihood estimator for $\boldsymbol{\theta}$ is defined as

$$\begin{aligned}\boldsymbol{\theta}_{ML} &= \arg \max_{\boldsymbol{\theta}} p_{model}(\mathbb{X}; \boldsymbol{\theta}) \\ &= \arg \max_{\boldsymbol{\theta}} \prod_{i=1}^m p_{model}(\mathbf{x}^{(i)}; \boldsymbol{\theta})\end{aligned}\tag{3.1}$$

To avoid taking a product over many values, which could lead to numerical underflow, Equation 3.6 can be also formulated in terms of a sum of logarithms, without affecting the arg max value

$$\boldsymbol{\theta}_{ML} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^m \log p_{model}(\mathbf{x}^{(i)}; \boldsymbol{\theta})\tag{3.2}$$

due to the above, this can be expressed as an expectation with respect to the empirical distribution \hat{p}_{data} described by the training data

$$\boldsymbol{\theta}_{ML} = \arg \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x} \sim \hat{p}_{data}} \log p_{model}(\mathbf{x}; \boldsymbol{\theta})\tag{3.3}$$

MLE can also be described in the context of minimising the dissimilarity between the model distribution p_{model} and the empirical distribution \hat{p}_{data} , measured by the Kullback-Leibler divergence:

$$D_{KL}(\hat{p}_{data} \mid p_{model}) = \mathbb{E}_{\mathbf{x} \sim \hat{p}_{data}} [\log \hat{p}_{data}(\mathbf{x}) - \log p_{model}(\mathbf{x})]\tag{3.4}$$

As the left term of the above equation is a function of only the training data, when training the model, only the right term requires minimisation:

$$- \mathbb{E}_{\mathbf{x} \sim \hat{p}_{data}} [\log p_{model}(\mathbf{x})]\tag{3.5}$$

This is equivalent to the likelihood term being maximised in equation 3.2, showing that maximising the likelihood corresponds to minimising the negative log likelihood, or KL divergence, between the model and training data distributions. This, in turn, corresponds to maximising the *cross-entropy* between these distributions.

As the number of training examples nears infinity, the maximum likelihood estimates of parameters approach their true value, provided that the following conditions are met:

- p_{data} must lie within the model family $p_{model}(\cdot; \theta)$
- p_{data} must correspond to a single value of θ

A detailed overview of MLE in the context of machine learning is given by Bengio, Goodfellow and Courville (2017).

3.2.4 Types of learning algorithms

No single best ML algorithm exists and, depending on task, available data and computational resources, many algorithms are applicable. ML algorithms are commonly categorised based on presence and type of supervisory signal:

- supervised learning
- unsupervised learning

Supervised Learning

Cat classification, as described in Section 3.2.2, is an example of a supervised learning problem, since examples of both the input image and desired output are provided to the model at training time. As such, a supervised learning problem aims to approximate the function $g : \mathbf{X} \mapsto \mathbf{Y}$ mapping input images \mathbf{X} to labels \mathbf{Y} . This is commonly accomplished by estimating a conditional probability distribution $P(y | x)$. Maximum likelihood estimation (MLE) is used to find the parameter vector θ which maximises the likelihood of observing the training labels Y given images X and a parametric family of distributions $P(Y | X; \theta)$. Accordingly, the conditional maximum likelihood estimator, given training data and labels is:

$$\theta_{ML} = \arg \max_{\theta} P(\mathbf{Y} | \mathbf{X}; \theta) \quad (3.6)$$

If the training examples are assumed to be independent and identically distributed (i.i.d) random variables, then Equation 3.6 can be decomposed into:

$$\theta_{ML} = \arg \max_{\theta} \sum_{i=1}^m \log P(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}; \theta) \quad (3.7)$$

This illustrates the intuition behind supervised ML techniques: given some training data consisting of m examples $\mathbb{X} = \{x^{(1)}, \dots, x^{(m)}\}$ and corresponding labels, and a parametric family of distributions (the model), the process of training consists of finding a set of parameters θ , which maximise the log likelihood of the labels given the input data.

Unsupervised Learning

Unsupervised learning algorithms do not require labelled data in order to learn. Instead, they aim to identify underlying patterns in the training data. If supervised learning can be viewed as approximating a conditional probability distribution, unsupervised learning aims to approximate an a priori distribution $p(x)$ – the data-generating distribution. For example, in dimensionality reduction, the goal is to find a compact representation of the training data, which maximises the preserved information, while reducing the number of features required to represent it. One method of performing dimensionality reduction is by using Principal Component Analysis (PCA), which can be interpreted in the context of MLE. Given a set of n -dimensional vectors X and the desired dimensionality m the objective is to find an $m \times n$ orthogonal projection matrix A , which minimises the squared reconstruction error, defined as:

$$error_{reconstruction} = \arg \min_A \|X - A^{-1}AX\|^2 \quad (3.8)$$

Dimensionality reduction can also be seen as a subset of representation learning, a task often performed using unsupervised methods. Representation learning aims to learn and extract task-relevant features, while also commonly reducing the dimensionality of the data. Autoencoders are a popular tool for representation learning. They consist of two key elements - an encoder and decoder. The encoder extracts features from the input data, while the decoder tries to reconstruct the input data from these features. Since the task only requires unlabelled input data, which it tries to reconstruct, no labelling is required. In a sense, the input data is simultaneously used as the ground truth labels and the autoencoder learns by minimising some reconstruction error between the input and output. Once the error is sufficiently low, the encoder can then be used to extract task-relevant features from input and perform further tasks using these features.

3.2.5 Gradient-based Optimisation

Many ML algorithms discussed in this work rely on the process of optimisation, which is commonly used for computing maximum likelihood estimates of model parameters. Optimisation aims to minimise or maximise the value of some function $f(\mathbf{x})$ by making changes to \mathbf{x} . In the context of optimisation, this function is often referred to as the *objective function* or the *criterion*. In scenarios where this function is specifically minimised, such as DNNs, it can be referred to as the *cost function*, *loss function* or *error function*. In practice, optimisation is commonly carried out using **gradient descent**

(Cauchy, 1847). For a real-valued function of a real variable $y = f(x)$, gradient descent makes small changes to x in order to minimise y . It accomplishes this by calculating the derivative of $f(x)$ with respect to x , which indicates the slope of $f(x)$ at point x . Using the sign of the resulting derivative, y can be incrementally minimised by changing x in the direction opposite to the sign of the derivative:

$$x_i := x_i + \Delta x_i \quad (3.9)$$

where

$$\Delta x_i = -\alpha \frac{df(x)}{dx_i} \quad (3.10)$$

Here, α is the size of the step taken, often referred to as the *learning rate* in machine learning problems.

This concept can be easily extended to functions with multiple inputs by using partial derivatives. If \mathbf{x} is a length n vector, then the gradient is defined as:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \quad (3.11)$$

For functions which output vectors, as well as taking vectors as input, the concept of the gradient can be extended to the **Jacobian** matrix. For a function $\mathbf{f} : \mathbb{R}^m \mapsto \mathbb{R}^n$, the Jacobian matrix $\mathbf{J} \in \mathbb{R}^{n \times m}$ of function \mathbf{f} is defined as:

$$J_{i,j} = \frac{\partial f(\mathbf{x})_i}{\partial x_j} \quad (3.12)$$

3.2.6 Stochastic Gradient Descent

Many ML problems require large training datasets in order to achieve good generalisation. For some task domains, such as image classification, computing a Jacobian matrix and performing a single gradient descent step a large dataset may become computationally expensive. Stochastic gradient descent (SGD) is an adaptation of gradient descent which addresses this problem. SGD simply computes the gradient for small random samples (commonly between 2 and 512) of the training dataset, referred to as *minibatches*. This gradient estimate is then used to perform a step of gradient descent, and the process is repeated for another minibatch. This can be illustrated as follows. If $J(\boldsymbol{\theta})$ is the negative conditional log-likelihood of the training dataset:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}, y \sim \hat{p}_{data}} L(\mathbf{x}, y, \boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^m L(\mathbf{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}) \quad (3.13)$$

where L is the loss term

$$L(\mathbf{x}, y, \boldsymbol{\theta}) = -\log p(y \mid \mathbf{x}; \boldsymbol{\theta}) \quad (3.14)$$

then the gradient of the cost function with respect to the model parameters and training dataset of size m is

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^m \nabla_{\boldsymbol{\theta}} L(\mathbf{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}) \quad (3.15)$$

.

Accordingly, the estimate of the gradient calculated on a minibatch $\mathbb{B} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m')}\}$ containing m' training examples corresponds to

$$\mathbf{g} = \frac{1}{m'} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m'} L(\mathbf{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}) \quad (3.16)$$

This estimate is then used to update the parameters

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \mathbf{g} \quad (3.17)$$

where ϵ is the *learning rate* parameter controlling the size of the step taken along the negative gradient.

SGD is among the most popular optimisation algorithms used in ML and particularly within deep learning, where many variants and extensions to SGD have been proposed. See Ruder (2016) for a detailed overview. Examples of these include SGD with momentum (Qian, 1999), which performs gradient updates based on an average of gradients at previous time-steps; Adam - a popular extension of SGD proposed by Kingma and Ba (2014), which incorporates per-parameter learning rates. The implementation of SGD in deep feedforward neural networks (discussed in Section 3.3) relies on the backpropagation algorithm (Rumelhart, Hinton and Williams, 1986).

3.2.7 Generalisation

In order for machine learning models to be useful, they must generalise well. Generalisation refers to the ability of the model to perform well on data which it has not been trained on. A common method of measuring model generalisation involves comparing the training

and test error. These are a result of the performance metric P evaluated on the training set and a hold-out test dataset, respectively. The test set contains examples from the same task domain, but not used during training. Bengio, Goodfellow and Courville (2017) suggest that model performance is determined by two key factors:

- training error
- difference between training and test error

A low training error indicates successful training, while a low test error suggests good generalisation. The scenario where a model cannot achieve sufficiently low training error is referred to as *underfitting*. Conversely, *overfitting* occurs when training error is low, but test error remains relatively high. Underfitting and overfitting are common practical challenges encountered in machine learning and are commonly a function of model *complexity* (*inf.* model capacity) and, in some cases, training dataset size and distribution. Model capacity refers to the ability of a given machine learning algorithm to approximate a large range of functions. The range of available functions, and thus capacity, can be changed by controlling the model’s *hypothesis space* (Bengio, Goodfellow and Courville, 2017). E.g. a linear regression model’s hypothesis space includes all linear functions $y = ax + b$. In practice, how model capacity can be adjusted depends on the type of ML algorithm used. E.g. in deep neural networks, model complexity can be increased by making the network “deeper” through adding layers of neurons, for example.

3.2.8 No Free Lunch Theorem & Regularisation

In his seminal work, Wolpert and Macready (1997) show that no single ML algorithm can outperform others, when averaged across all data-generating distributions and tasks. Dubbed as the *no free lunch theorem*, this finding indicates that for particular task-related distributions, appropriate ML algorithms must be selected. Thus, as Bengio, Goodfellow and Courville (2017) suggest, design of effective ML models relies on understanding the task-relevant data-generating distributions and selecting an ML algorithm that performs well on data sampled from such distributions. Moreover, task-specific design allows for incorporation of constraints, or prior knowledge, which in turn, can allow for performance gains. Effectively, such constraints can be used to limit the hypothesis space and impose a preference for certain solutions, compared to others. The process of introducing such preferences is referred to as *regularisation* and is key to the ability of many ML algorithms to generalise well to unseen data, despite learning from only a subset of the true data-generating distribution.

Regularisation has been shown to reduce overfitting and improve generalisation of ML models Krogh and Hertz (1992). A technique known as **weight decay** is a simple, yet effective regularisation method, which expresses preference for lower-valued model parameters. Weight decay is implemented using a **regulariser** Ω - a penalty, consisting of the $L1$ or $L2$ norm of all the model’s weights (parameters).

$$\Omega(\mathbf{w}) = \mathbf{w}^\top \mathbf{w} \quad (3.18)$$

This regulariser is added to the training loss, for example MSE between predictions and ground truth labels, which forces the learned weights to take on smaller values. The strength of this regularisation can be controlled by a scalar λ :

$$J(\mathbf{w}) = MSE_{train} + \lambda \Omega(\mathbf{w}) \quad (3.19)$$

Similarly to the choice of ML algorithms for a given problem, no single best regularisation method exists and many implicit and explicit approaches to regularisation have been proposed in the literature. Some of these are discussed later, in the context of specific ML problems addressed in this thesis.

3.2.9 Hyperparameters

In addition to the parameters of a machine learning model estimated during the training process (the weights), a set of parameters controlling the dynamics of the learning algorithm itself must be specified. These are referred to as **hyperparameters** and are specific to the particular model used for a given task. The λ parameter controlling the strength of weight decay described above is an example of a hyperparameter. A common reason for why hyperparameters are not inferred during the training process is that this leads to overfitting. For example, allowing λ to be determined by the training process in the regularisation example above would result in $\lambda = 0.0$, since this minimises the training error. Any other parameter influencing model capacity will have the same effect, as increasing complexity is the “easiest” way for the model to lower its training error.

Several approaches have been proposed to find optimal hyperparameter values. These are commonly based on a process consisting of sampling a set of hyperparameter values, training and evaluating a model and repeating the process, until the training or generalisation error is minimised. The sampling strategy adopted can be based on a random or grid-based search Bergstra and Bengio (2012), evolutionary algorithms Young et al. (2015) or Bayesian methods Klein et al. (2016), to name a few.

3.2.10 The essence of a machine learning algorithm

Most machine learning algorithms can be described in the context of four key components: a training dataset \mathbb{X} , a cost function $J(\boldsymbol{\theta})$, an optimisation procedure (e.g. SGD) and a parametric model family p_{model} . Thus, a large range of possible ML algorithms can be designed by replacing any of these components, depending on the requirements of a given task. As described in Section 3.2.3, for p_{model} to accurately approximate \hat{p}_{data} , \hat{p}_{data} must lie in the model family $p_{model}(\cdot; \boldsymbol{\theta})$. Thus, the model’s ability to approximate a wide range of complex functions is fundamental to solving non-trivial ML problems. In recent

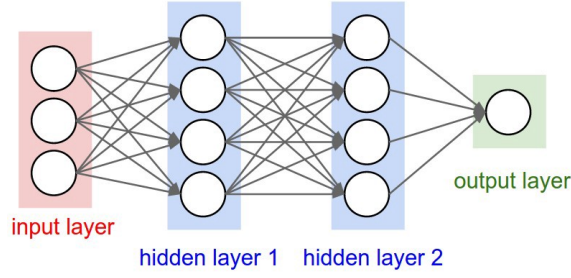


Figure 3.1: Conceptual illustration of a neural network, consisting of an input layer, two hidden layers and an output layer. Each node represents a neuron. The arrows represent forward connections between neurons.

years, **Feedforward neural networks** (or **multilayer perceptrons**) have become the de-facto standard in practical applications of machine learning. The term *deep learning* refers to using such deep, hierarchical models in combination with optimisation methods, such as SGD, discussed earlier in this chapter, to solve difficult problem, such as object classification, speech recognition or image synthesis. Deep learning is a vast and rapidly-developing research domain, while key concepts are discussed below, the interested reader is referred to Bengio, Goodfellow and Courville (2017) for an in-depth review.

3.3 Deep Feedforward Networks

Deep feedforward neural networks, commonly referred to as deep artificial neural networks (DNNs), are general-purpose function approximators. A trained DNN approximates some function f^* , for example, for a classification task, the network defines a mapping $\mathbf{y} = f(\mathbf{x}; \boldsymbol{\theta})$ and uses SGD in order to find the set of parameters $\boldsymbol{\theta}$ resulting in the best approximation, as measured by some cost function $J(\boldsymbol{\theta})$. A DNN can be seen as a composition of multiple functions, e.g. $f(\mathbf{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$. Each of these functions is referred to as a **layer** of the neural network, and the depth of a DNN is given by the number of these consecutively chained layers. The final layer of a DNN is referred to as the output layer. During training, the cost function is computed using values output by this layer. All other layers in the network, bar the input layer, are referred to as **hidden layers**, as their desired outputs are not explicitly represented in the training data.

Each layer in the network consists of multiple **units**, also known as neurons. Each neuron computes a function of its inputs, commonly a vector of all outputs of the previous layer, and outputs a single activation value.

3.3.1 The Neuron

A unit, or artificial neuron, builds on the concept of a linear function of vector-valued input, by introducing a nonlinearity. In the DNN context, a neuron consists of a set of inputs $\mathbf{x} \in \mathbb{R}^m$, corresponding *weights* $\mathbf{w} \in \mathbb{R}^m$ and a nonlinear function h applied to the sum of the inputs \mathbf{x} scaled by weights \mathbf{w} :

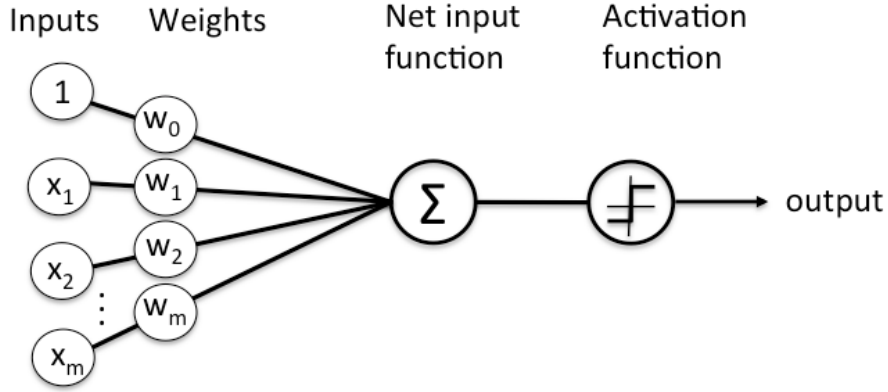


Figure 3.2: Illustration of an artificial neuron, consisting of inputs x_0 through x_m , corresponding weights w_0 through w_m , summing Σ and activation function.

$$a = h(z) \quad (3.20)$$

where

$$z = \sum_{i=0}^m w_i x_i \quad (3.21)$$

Here, a represents the real-valued activation of the neuron. Commonly, x_0 is assigned value +1 and represents the *bias* term. This can also be represented as an explicit additive term.

The set of weights \mathbf{w} of each neuron constitutes the parameters $\boldsymbol{\theta}$ learned by the network during training. While many neural network implementations extend or adapt this definition of the neuron, the guiding principles remain the same: find the parameter values $\boldsymbol{\theta}$ that minimise cost function $J(\boldsymbol{\theta})$

3.3.2 Activation Function

The activation function h is key to the success of DNNs in many practical applications. Without a nonlinear activation function, the output of a single neuron would remain a linear function of its inputs. A network, constructed of such neurons, would thus also compute a linear function of its inputs. Activation functions introduce nonlinearity into the output of neurons, thus allowing a neural network to learn a nonlinear function of its inputs. This also allows for chaining of multiple layers of the network, and thus expressing complex features as a function of simpler features.

The practical challenges associated with training DNNs have driven research into different activation functions. Aspects such as efficiency, continuous differentiability, range and smoothness are all desired properties.

Rectified Linear Units

A standard activation function used in a range of state-of-the-art DNNs is the rectified linear unit (ReLU) (Nair and Hinton, 2010):

$$h(z) = \max\{0, z\} \quad (3.22)$$

where z is the sum of inputs \mathbf{x} weighted by \mathbf{w} , as in Equation 3.21. For $z > 0$ the function outputs the input value, otherwise it outputs 0. Its desirable properties include computational efficiency and constant gradient of 1 for $z > 0$, however it suffers from the *dying ReLU problem*, due to the fact that the gradient is 0 when $z = 0$. In those scenarios, gradient descent will no longer affect the associated weights, thus removing the contribution of that neuron to the task.

ELU, Leaky ReLU, SeLU

Many extensions of the ReLU function attempt to address the dying ReLU issue, these include the exponential linear unit (ELU) (Clevert, Unterthiner and Hochreiter, 2015):

$$h(z) = \begin{cases} z & \text{if } z > 0 \\ \alpha(e^z - 1) & \text{otherwise} \end{cases} \quad (3.23)$$

where α is a hyperparameter to be tuned. The ELU unit behaves as an identity function for $z > 0$ and smoothly approaches $-\alpha$ for $z \leq 0$.

The leaky ReLU (Maas, Hannun and Ng, 2013) is another approach to addressing the dying ReLU problem. Instead of outputting 0 when $z \leq 0$, leaky ReLU retains a small slope, controlled by parameter α :

$$h(z) = \begin{cases} z & \text{if } z > 0 \\ \alpha z & \text{otherwise} \end{cases} \quad (3.24)$$

where α is usually set to a small value, e.g. 0.01.

Another example of rectifier-based activation functions is the *scaled exponential linear unit* (SeLU) (Klambauer et al., 2017):

$$h(z) = \lambda \begin{cases} z & \text{if } z > 0 \\ \alpha e^z - \alpha & \text{otherwise} \end{cases} \quad (3.25)$$

Logistic Sigmoid & Hyperbolic Tangent

The logistic sigmoid

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \quad (3.26)$$

and hyperbolic tangent

$$\tanh(z) = 2\sigma(2z) \quad (3.27)$$

are examples of activation functions whose output value ranges are bounded. This is useful for some tasks where the output values must be within some specified range, however due to the property of saturation (the output of a function being close to 0 for small values and close to 1 for high values), models with these activation functions in hidden layers are difficult to train using gradient-based methods. The logistic sigmoid is, however, often used as the output layer of networks trained to perform binary classification.

Softmax

The softmax activation function serves an important practical purpose of transforming an input vector into a probability distribution. It is commonly used in the final layer of a classification neural network, where the number of output neurons K corresponds to the number of classes. Accordingly, the softmax activation converts the linear activation z_k of the k^{th} neuron in the OUTPUT layer into a probability score corresponding to the k^{th} class. Accordingly, the softmax activations for all classes sum to 1.0. The softmax function is defined as:

$$softmax(z_k) = \frac{\exp(z_k)}{\sum_{j=1}^K \exp(z_j)} \quad (3.28)$$

where z_k is the linear activation of the k^{th} unit, as defined in Equation 3.21.

Linear

The linear activation function is equivalent to not specifying any activation function and outputting z . Linear activation functions are commonly used in the final layer of neural networks tasked with regression problems, which estimate the mean of a conditional Gaussian distribution. In such scenarios, the desired output range may be unbounded.

3.3.3 Cost Functions

For a given ML task, the choice of cost function has a large impact on the results. This is because the gradient of the cost function directly influences the changes to the model's

weights, as shown in Equations 3.13 - 3.17. In practice, cost functions are commonly selected based on the task being performed by the network. In classification tasks, the cost function commonly involves a cross-entropy between the target \mathbf{y} and predicted $\hat{\mathbf{y}}$ multinomial distributions:

$$CE(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_i^C y_i \log(\hat{y}_i) \quad (3.29)$$

In regression tasks, the cost function measures the mean squared error between ground truth and predicted values (or vectors thereof):

$$MSE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{m} \sum_i^m (y_i - \hat{y}_i)^2 \quad (3.30)$$

, while in representation learning, the reconstruction loss commonly calculates the mean absolute error between predictions and ground truth.

$$MAE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{m} \sum_i^m |y_i - \hat{y}_i| \quad (3.31)$$

Many adaptations and extensions of these cost functions have been proposed based on task type and the structure of inputs and outputs. These extensions commonly aim to address issues such as class imbalance in the training dataset, or allow expressing preference between precision and recall in classification tasks (Lin et al., 2017). Special cases of cost functions will be discussed later, in the context of individual chapters.

3.3.4 The Universal Approximation Theorem

The practical success of DNNs is largely associated with their theoretical ability to approximate any given function. The universal approximation theorem for neural networks (Hornik et al., 1989) states:

“There is a single hidden layer feedforward network that approximates any measurable function to any desired degree of accuracy on some compact set K .”

While this does not guarantee that such function may be practically learnable using current gradient-based learning algorithms, it shows that a sufficiently large neural network is capable of representing any function.

3.3.5 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a special case adaptation of conventional neural networks, as discussed in Section 3.3, applied to processing 2- or 3-dimensional data,

such as images. From a perspective of implementation, CNNs simply replace conventional matrix multiplication operations with convolution, in at least one layer:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n) \quad (3.32)$$

Where $S(i, j)$ is the result of the convolution operation, also represented by $*$, I is the input image and K is the convolution kernel. For the purposes of CNNs cross-correlation is typically used instead of convolution, which in this context is equivalent, as it only involves flipping the convolutional kernel:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n) \quad (3.33)$$

In practice, CNNs are well suited to the task of processing image data, due to the fact that they exploit three key ideas, summarised by Bengio, Goodfellow and Courville (2017) and discussed below.

Sparse Interactions

In conventional, densely-connected DNNs each input value interacts with each output value of the network, in CNNs the use of the convolution operation results in the information at the output being a function of a local neighbourhood of the input. In practice, this greatly reduces the number of operations, speeding up training and evaluation of such networks.

Parameter Sharing

While in DNNs each individual input to a layer of the network is associated with a single parameter, CNNs share parameters across inputs. For example, a given layer of a CNN may consist of D $k \times k$ filters with learnable weights, allowing it to learn D different $k \times k$ convolutional kernels, outputting D feature maps, each resulting from convolving the input image with one of the D kernels. Here, the number of parameters per layer is a function of kernel size and count, but not a function of input dimensionality, due to the sharing of each kernel across the entire input array.

Equivariant Representations

CNNs generate translation-equivariant representations. This means that a translation applied to the input image pixels will result in an equivalent translation of the corresponding features in the resulting feature map. This is a useful property of convolutions when applied to images, as it allows for parameter sharing and localisation of features.

3.3.6 Properties of Convolutional Layers

Several properties of convolution operations, as applied in CNNs, can be adjusted in order to influence the behaviour of the operation. These include:

- **kernel size:** the dimensions of the convolutional kernel, e.g. 3×3
- **stride:** the size of the step taken by the kernel
- **padding:** the amount of zero-padding added to the input image
- **number of filters (kernels):** controls the number of kernels learned by a convolutional layer and thus the number of different features extracted.

These properties constitute some *hyperparameters* of CNNs are key in controlling their model capacity, scale of features at different layers and other task-dependant properties. The settings of these hyperparameters are often evaluated experimentally, with respect to task, model architecture and dataset, see Bergstra and Bengio (2012) for an overview.

3.3.7 Other Neural Network Layers

Various specialised layers and aspects of network architectures have been proposed in order to improve performance, generalisation and training dynamics of DNNs.

Batch Normalisation

Batch normalisation (BN) (Ioffe and Szegedy, 2015) is a process aiming to standardise the mean and standard deviation of a layer’s output activations (and consequently the following layer’s input activations) for a given minibatch. This speeds up training through reducing the change in the distribution of inputs to successive layers, as the weights of the network are adjusted over time. This phenomenon is known as *internal covariate shift*. Despite some disagreement regarding the theoretical grounds for the effectiveness of BN (Kohler et al., 2018), experimental evidence has made it a standard component of modern neural networks.

Dropout and Spatial Dropout

Dropout (Srivastava et al., 2014) is an efficient regularisation method commonly used in DNNs. Dropout works by randomly ignoring the contributions of certain neurons and their associated connections to the network’s output during training. This approximates the process of training many parallel subnetworks to perform the same task and pooling their outputs, reducing strong correlations between activations of a network. Another interpretation of dropout is that it adds noise to the training process, by effectively ignoring a certain proportion of activations randomly in each iteration. This forces different subsets of the network to perform well on the training task, and thus acts as a regulariser. In CNNs, due to the convolution operation, dropping individual pixels from feature maps

has little effect on the output, rendering dropout ineffective. To address this, Tompson et al. (2015) proposed a spatial adaptation of dropout. *Spatial dropout* works by dropping entire feature maps, instead of dropping individual neurons (which correspond to pixels in the feature maps).

3.4 Architectures & Applications of CNNs

CNNs are among the most popular classes of neural networks. They have been successfully applied to many pattern recognition problems and have contributed to a revolution in modern computer vision, allowing significant progress in many long-standing problems. As no standardised, formal approach to designing DNN architectures exists, many solutions are available, and a complete review is beyond the scope of this thesis. This section discusses key CNN-based approaches to solving a range of vision-related tasks, the associated network architectures, cost functions and other contributions relevant to this thesis. For the sake of organisation, the work discussed here is organised by task domain, however it is important to note that due to the general-purpose nature of NN-based approaches, these techniques are relevant in the context of the aim of this thesis.

3.4.1 Image Classification

The first broadly successful application of a CNN architecture to a large-scale visual problem was carried out by Krizhevsky, Sutskever and Hinton (2012). They proposed a 5-layer convolutional neural network (dubbed AlexNet), which achieved near-human performance in image classification on the challenging ImageNet Large-Scale Visual Recognition Challenge dataset consisting of 1000 object classes and 1.2 million training images. This seminal work commenced the modern deep learning era in computer vision. Consequently, DCNNs have been applied to a host of visual tasks where input images are mapped to some output representation, label or image. The ILSVRC continued to attract CNN-based solutions, seeing adaptations of AlexNet, such as Clarifai (Zeiler and Fergus, 2014), or VGG16 (Simonyan and Zisserman, 2014) win the challenge in the following years. In 2014, GoogLeNet (a.k.a. Inception V1), proposed by Szegedy et al. (2015), took an entirely different approach by proposing *inception modules* which perform parallel convolution at different scales and concatenate the result. In addition to this, the authors improved the efficiency of various aspects of the network, allowing them to drastically increase its depth to 22 layers. This was surpassed the following year by He et al. (2016) who proposed networks based on residual blocks, which addressed many practical problems associated with training very deep networks, such as the vanishing gradient problem (Hochreiter, 1998) and the degradation problem. Despite continued development of new architectures, these models have remained a key building block, with many researchers adopting them as a starting point for transfer learning or modifying and extending them to new tasks.

3.4.2 Object Detection

CNNs for object classification were soon adapted to object detection tasks, allowing for detection and classification of multiple objects in a single image. An object detector returns the class and location of each detected object instance. The location is commonly described as a set of bounding box coordinates. Early contributions were made by Girshick et al. (2014) and Girshick (2015) through applying a conventional CNN classifier to multiple pre-specified regions of an input image. This was further extended by Ren et al. (2015), who trained a specific region proposal network to replace pre-computed sets of regions. Redmon et al. (2016) proposed a novel single-stage detector (SSD), which combined the region proposal and object classification stages into an end-to-end network, capable of performing in real-time. Multiple improvements of this model were later suggested (Redmon and Farhadi, 2017, 2018). The SSD object detection framework has become fundamental in real-world applications of computer vision, such as autonomous driving (Chen et al., 2016), traffic monitoring (Lin and Sun, 2018), security (Akçay et al., 2018) or medicine (Sarıkaya, Corso and Guru, 2017). A detailed review on CNN-based object detection is given by Zhao et al. (2019).

3.4.3 Semantic Segmentation

In the context of CNNs, the task of semantic segmentation is closely related to object detection and classification. However, instead of outputting object classes and instance bounding boxes, semantic segmentation produces a pixel-wise class map, assigning an object class to every pixel of the image, thus performing segmentation in addition to detection and classification. The problem of semantic segmentation fuelled the development of fully convolutional networks (FCNs) (Long, Shelhamer and Darrell, 2015). As opposed to conventional CNNs, which output a single class label, or bounding box coordinates, these networks output images, in addition to taking them as input. Such networks are well-suited to image-to-image translation problems and have been used in areas such as image colourisation (Cheng, Yang and Sheng, 2015; Zhang, Isola and Efros, 2016), denoising (Xie, Xu and Chen, 2012), style transfer (Gatys, Ecker and Bethge, 2016) and super-resolution (Johnson, Alahi and Fei-Fei, 2016). An attractive property of many of these applications stems from the fact that labelled training data can be generated on the fly. For example, in the context of denoising, the noisy input images can be generated on-the-fly, by applying various amounts of noise to clean images. The *Unet* is a popular example of a general-purpose FCN architecture developed to address semantic segmentation (Ronneberger, Fischer and Brox, 2015) and applied to a wide range of other image-to-image problems, notably as the generator in Pix2Pix, a popular generative adversarial network architecture for image synthesis and translation (Isola et al., 2017). The *Unet* is based on an encoder-decoder architecture, crucially adding *skip connections* which preserve low-level features and help with gradient backpropagation (Drozdal et al., 2016).

3.4.4 Image Synthesis

The development of FCNs, particularly Unet-style image generation networks, coupled with the framework of a Generative Adversarial Network (GAN) (Goodfellow et al., 2014) and their conditional counterpart (Mirza and Osindero, 2014) contributed to major advances in image synthesis. The aim of a GAN is to generate new data with the same properties as the training data. Thus, a GAN implicitly approximates the data-generating distribution through training two subnetworks, the generator (G) and discriminator (D) in an adversarial manner. G aims to generate samples that fool D, while D tries to discriminate between real samples from generated ones. Many state-of-the-art image synthesis models have adopted and further developed the GAN framework for purposes such as image super-resolution (Wang et al., 2018b), image translation (Zhu et al., 2017), inverse rendering (Thies et al., 2016a) and many other non-vision-based applications.

3.4.5 Perceptually-based Tasks

Crucially to the theme of this thesis, CNNs have also been used extensively for explicit modelling of aspects of human perception. Such approaches are often based on training directly on subjective data, commonly collected in task-specific experiments with human observers. As illustrated throughout this chapter, a key benefit of DL and gradient-based optimisation techniques is that they can be adapted to new tasks with relative ease, provided sufficient training data are available. In practice, this may also involve adaptations to the architecture of the CNN to account for the task-specific properties, such as dimensionality of the inputs and outputs, model capacity or regularisation strategies employed.

Saliency Prediction

Saliency prediction, the goal of which is to accurately predict the probability distribution of visual attention in an image, is one example of such task. This task has recently been approached through direct learning of the perceptual function mapping input images to ground truth saliency maps based on empirical fixation distributions (Zhao et al., 2015; Pan et al., 2016). See Figure 3.3 for an example. Other approaches have also proposed architectural improvements, such as extraction of features at multiple scales (Cornia et al., 2016), or extensions for predicting saliency in video sequences (Jiang et al., 2018).

Perceptual Similarity

Another example of using empirical subjective data in order to generalise human performance to new stimuli has been put forward by Zhang et al. (2018b), who trained a perceptual similarity metric based on subjective similarity scores, collected for a large dataset of image patches using a two-alternative forced choice (2AFC) procedure. A CNN was then trained to map pairs of input patches to a single-value perceptual similarity scores. The authors found the proposed perceptual metrics to outperform classical metrics,

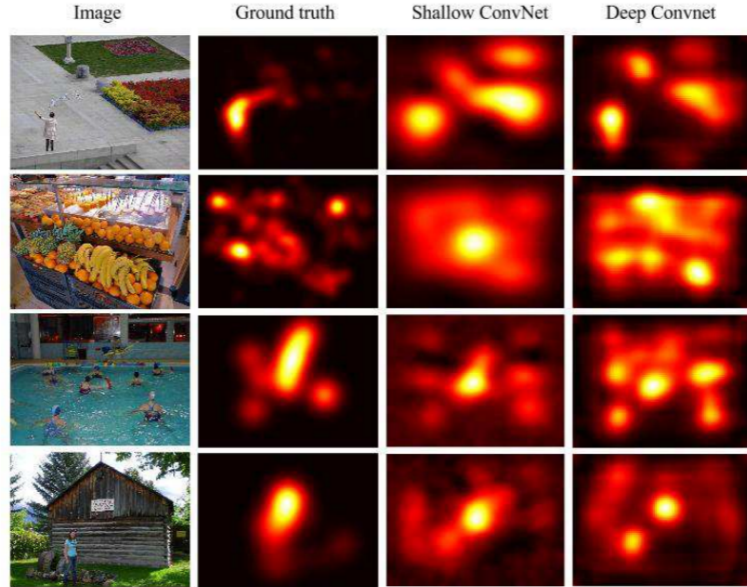


Figure 3.3: Examples of empirical saliency maps compared to saliency maps generated by a shallow and deep CNN. Image courtesy of Pan et al. (2016)

based on analysis on benchmark datasets. This approach has recently been applied in the context of audio similarity (Manocha et al., 2020), achieving state-of-the-art results.

Image Quality Assessment

Applications of CNN-based models can also be found in image quality assessment. Most of these approaches build on the premise of learning mappings between images and corresponding subjective quality scores (Gu et al., 2014). Bosse et al. (2017) evaluate such an approach in the context of both NR and FR IQA, Hou et al. (2014) replace numeric scores with linguistic descriptors during data collection and training, while Kim and Lee (2017) propose a model that also learns visual error sensitivity maps from FR image quality datasets. Bianco et al. (2018) report on a broad study of architectural choices for DL-based NR IQA, concluding that patch sampling and average pooling are both reliable design choices for this application. Based on the insight that group IQA scores tend to be distributed, rather than concentrated, Talebi and Milanfar (2018) proposed a training regime based on representing subjective scores as distributions, rather than integers. As public dataset sizes increase (Hosu et al., 2020), this task is likely to continue receiving attention in the research community. For an extensive survey of recent perceptually-based IQA solutions, see Zhai and Min (2020). Overall, the success of DL-based IQA solutions illustrates that under appropriately controlled conditions, deep learning methods are effective at approximating and generalising human performance in perceptual tasks.

Image Transformation & Harmonisation

CNNs have been extensively studied in the context of realistic image synthesis and transformation. In addition to synthesis of new image content, described earlier in this section, these techniques have been leveraged for conditional image transformation and synthesis. Similarly to the statistical approaches to composite harmonisation described in Section 2.7, the goal of these techniques is to transform an input image, often a composite, or CGI, for practical or aesthetic purposes. Applications of such techniques span a broader range of tasks, such as domain adaptation (Murez et al., 2018) or style transfer (Gatys, Ecker and Bethge, 2016). Thanks to the general nature of image-to-image DL architectures, this allows for learning of a wide range of functions mapping between images without major modifications to the network architecture (Goodfellow et al., 2014).

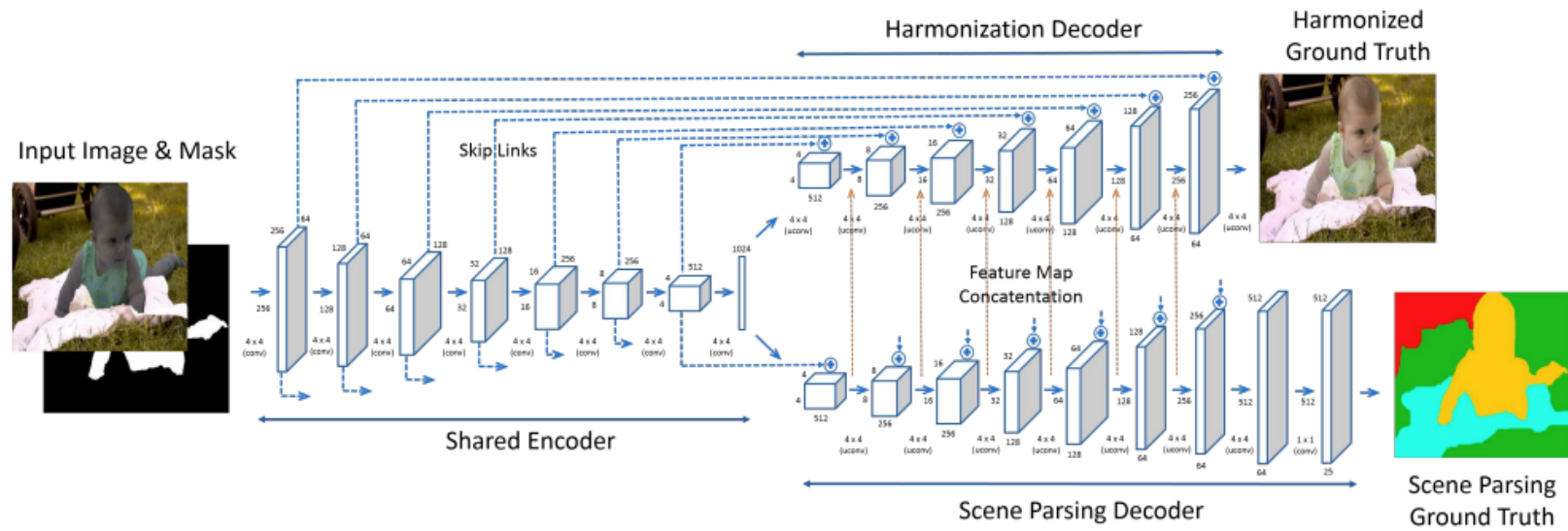


Figure 3.4: Architecture of the original DIH model. Image courtesy of Tsai et al. (2017)

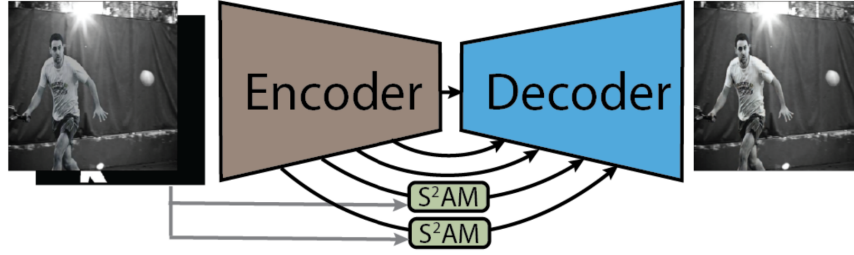


Figure 3.5: Illustration of the spatial-separated convolution module attached to a U-Net style, encoder-decoder architecture. Image courtesy of Cun and Pun (2020)

Image harmonisation has received much attention in this context. Tsai et al. (2017) propose an approach to harmonising an image composite, given the original, unharmonised image input and a binary mask indicating the foreground object. They also leverage features learned on a semantic segmentation task in order to improve spatial allocation of transformations, in essence learning a mapping between input and target composite images, conditioned explicitly on object masks and implicitly on scene semantics. The network is trained by minimising either L1 or L2 distance between the output image and ground truth. Figure 3.4 illustrates the architecture of the CNN used by the authors, which consists of a shared encoder and two separate decoders, one used for harmonisation of the input composite, the other for scene segmentation. The intermediate feature maps from this scene segmentation branch are concatenated with corresponding features in the harmonisation decoder branch. However, this method is difficult to extend to novel data, since it relies on input composite masks, as well as scene segmentation ground truth data to be available at training time, in addition to the unharmonised and harmonised images. The input object mask is also required at inference time, which makes this method impractical for legacy content (such as films or photographs, for which the original source masks may not be available), since it requires manual generation of a mask.

This general approach was extended by Cun and Pun (2020), who proposed a spatial-separated convolution module, designed to extract more relevant harmonisation features, by relying on the composited region for feature extraction. Using the input binary mask, intermediate feature maps are re-weighted according to the input mask and a trainable attention module (see Fig. 3.5). This architectural design allows for the network to treat the regions requiring harmonisation differently from the background, for example, to only manipulate the pixels in the target region, while influencing their appearance based on the properties of the background region. Similarly to other harmonisation networks of this design, this technique also requires input masks to be available, both at training and inference time.

The spatial-separated module was later adopted by Cong et al. (2020) who proposed a new standardised, extensive harmonisation dataset (iHarmony), and architectural improvements to the base model, including partial convolutions and an adversarial training

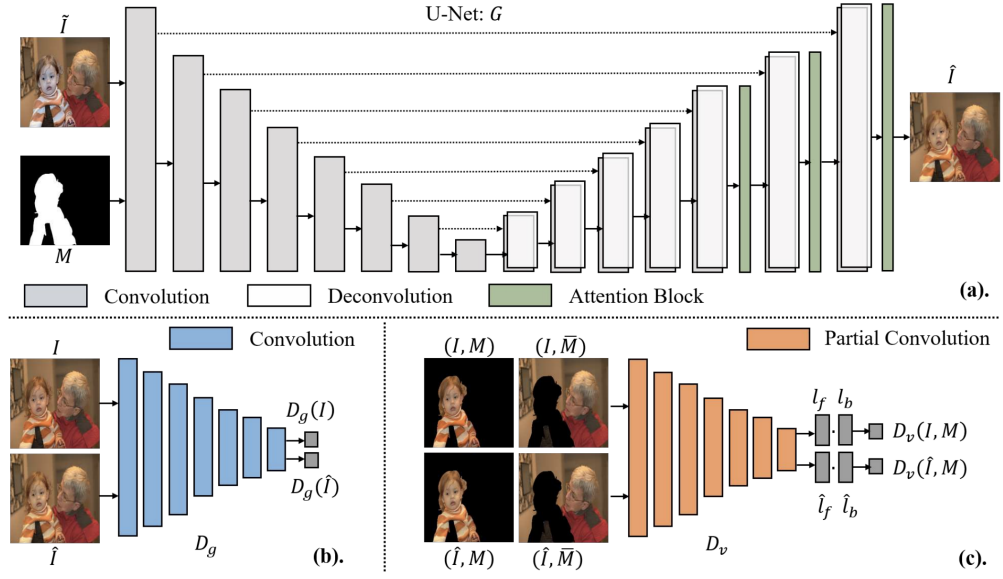


Figure 3.6: An illustration of the architecture of DoveNet, including the generator with attention modules from Cun and Pun (2020) and two discriminator networks used in the adversarial training procedure.

regime (see Fig. 3.6 for an illustration of this architecture). Here, partial convolutions allow for one of the discriminator networks to encode separate features for the background and foreground, allowing for a direct comparison of the appearance of the foreground and background, respectively, in the domain verification discriminator. In addition to this domain verification discriminator, a standard global discriminator is also used, and the losses are averaged. This method, despite its impressive results, is challenging to train, both due to the number of parameters in the network, and the volatile nature of adversarial training.

An alternative approach was proposed by Sofiuk, Popenova and Konushin (2020), who leverage neural networks pre-trained on semantic segmentation tasks in order to extract better appearance descriptors for local regions. The proposed approach does not leverage any characteristics of human perception, relying on a purely data-driven approach to harmonisation. A different approach is proposed by Chen and Kae (2019), who rely on a GAN-based architecture to perform the entire process of compositing, including the positioning and harmonisation of the foreground object in the target scene. While the approach achieves impressive results, it requires a complex adversarial training procedure and a custom synthetic CGI dataset, making this method challenging to build on in practical terms.

Visual Realism Assessment

As mentioned in Chapter 2, despite the broad adoption of ML techniques in image transformation and synthesis, few examples of using ML techniques in the context of visual

realism prediction exist. Zhu et al. (2015) design a CNN to perform binary classification of realism in images. This is accomplished by automatic generation of image composites, thus removing the requirement for human-annotated data, however only provides global predictions of realism, making it unsuitable for local prediction of realism. Similarly, Fan et al. (2018) use a CNN for binary prediction of realism based on an extensive subjective study evaluating impact of latent factors on perception of visual realism. The authors also propose an explainable model based on a support vector machine (Noble, 2006). While both techniques outperform prior approaches, they require extensive subjective modelling and some manual feature extraction. Furthermore, the approach also considers global image realism only, making it more suitable for CG vs photograph classification, but not for evaluation of local image composite quality. Both these approaches rely on a CNN to learn relevant features and classify the input. A different approach is taken by Yao et al. (2018), who instead rely on sensor pattern noise to distinguish CG from real images. The authors incorporate high-pass filtering into the network design, which performs well for global classification of manipulated vs authentic photographs. Furthermore, the high-pass filters require tuning for different datasets, making this method difficult to generalise to new domains.

3.5 Limitations of Existing Approaches

While many deep learning approaches to composite image harmonisation have been proposed, as illustrated in this chapter, they all suffer from common shortcomings. Firstly, as previously discussed, they commonly require additional input and/or output training data to be available during training and/or inference time. This makes it both challenging to adapt such methods to new datasets, but also difficult to apply to legacy image composites (e.g. ones for which masks are not readily available). Secondly, given that few composite datasets are publicly available, and their sizes are limited, training these methods on small novel datasets may yield results of lower quality, compared to the results presented on research datasets, making it increasingly difficult to adopt these techniques in practical settings. Thirdly, existing methods do not consider properties of human perception and adopt an exclusively data-driven approach, potentially reducing the resulting subjective quality, or realism. Furthermore, manually-generated object masks, while commonly available for new composites, do serve as a form of hard-coded prior knowledge, effectively side-stepping a crucial part of the human process of image compositing - assessing the type and magnitude of the mismatch, before the composite is harmonised. Explicitly incorporating detection of composite regions requiring correction in the harmonisation process would allow for closer replication of the process that humans undertake when performing image compositing. Additionally, this would also allow for tuning of the detection process based on human perceptual sensitivity, by effectively decoupling the detection from the correction of composite artefacts.

3.6 Summary

This chapter has reviewed the background and relevant literature in machine learning, including learning-based approximation of perceptual functions, such as those commonly modelled in image quality assessment. Due to their universality and broad range of practical applications, deep learning techniques, particularly ones based on CNNs, have been identified as a plausible approach to modelling complex functions, such as the perceptual functions mapping a visual stimulus to a subjective score. Despite the extensive body of work relying on CNNs to learn mappings between input images and subjective opinion scores, few attempts have been made towards directly approximating the function performed by observers under a visual realism assessment task and leveraging that function in downstream transformation tasks, such as harmonisation. To date, existing approaches to DL-based prediction of subjective image properties have been shown to rely on direct approximation of the realism discrimination function directly from image data, often performing the task globally, thus providing a single, image-wise assessment of realism, unsuitable for automatic improvement of composite realism. On the other hand, state-of-the-art harmonisation models achieve impressive results, commonly relying on pre-trained features from proxy tasks, such as semantic segmentation. This highlights the importance of the feature extraction stage for the success of both realism assessment and its improvement, and suggests that learning general-purpose feature representations based on the types of distortions present in image composites may be an effective route towards developing improved harmonisation and realism assessment models. The following chapters draw on these findings to design perceptually-based models of visual realism as a function of distortion visibility and generalise them using learning-based approaches.

Chapter 4

Modelling Perceptual Realism in Image Composites

This work was published in:

Dolhasz, A., Williams, I. and Frutos-Pascual, M., 2016. Measuring Observer Response to Object-Scene Disparity in Composites. *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE, pp.13–18

4.1 Introduction

The previous chapters have introduced the problems of image quality and visual realism, methods for their measurement and modelling, as well as various approaches to their prediction. Such subjective measures have been shown to be influenced by a range of visual features of the image stimuli, for example, resolution, sharpness, colour rendition or compression side effects, to name a few. As discussed, many objective IQA methods utilise some form of feature extraction or decomposition of the input image, commonly based on models of the HVS, or models of error sensitivity. The impact of variations in such features has been studied in the context of image quality assessment, however, similar studies are limited in the case of visual realism. In order to model visual realism as a function of particular image features, or distortions, it is paramount to first understand how subjective human judgments change, when such distortions are detected.

Building on the above observations, this chapter evaluates whether visual realism can be reliably modelled as a function of distortion perception and develops a psychophysical model of human sensitivity to a set of local distortions common to image composites. This is achieved through the use of *synthetic composites* - natural images with local transformations applied in order to simulate distortions common to image composites. Such synthetic composite images allow for individual distortions to be introduced in a controlled manner, making it possible to measure human response to incremental

changes. This also allows for mapping between the subjective responses and objective image differences, as both the *perfect* appearance and the type and magnitude of the introduced distortion are known *a priori*. Using these composite images, a two-alternative forced-choice study is then carried out to measure average group JNDs for different local composite distortions. These JNDs encode the amount of distortion required for observers to reliably distinguish between the synthetic composites and corresponding real images. Using approaches discussed in Section 2.3, psychometric functions are fit to the collected responses in order to model the relationship between the magnitude of local transformations and observer detection performance. The distortions introduced are based on similar studies and include exposure shifts, contrast scaling and colour temperature shifts. The resulting models are then evaluated and discussed in context of related work.

4.2 Background & Related Work

4.2.1 Importance of Visual Realism

Visual realism impacts systems well beyond aspects relating to quality or aesthetics. There is evidence that human task performance can be affected by the visual realism of a virtual environment or other task-specific image stimuli. This effect has been studied, particularly in the context of visual search (Lee et al., 2013; Ragan et al., 2015) as well as navigation in virtual spaces (Meijer, Geudeke and Van den Broek, 2009; Lokka et al., 2018). The performance increase noted in more realistic VEs is sometimes explained by the subjective increase in presence, which is often linked to task performance in VEs (Welch et al., 1996). It is important to mention, however, that higher visual realism does not always correlate with better task performance. Smallman and John (2005) argue that in many cases naive reliance on highly realistic visual displays can be detrimental and provide evidence based on geospatial data interpretation. This is corroborated in a subsequent study of visually realistic map renderings, which result in longer navigation task completion times and lower task accuracy compared to abstracted, less realistic line drawings of the same data (Wilkening and Fabrikant, 2011). Since visual realism is not a universal concept (Ferwerda, 2003) and can be modulated by both task and stimuli, the need for efficient approaches to its measurement and modelling is apparent. Understanding how visual realism is perceived by humans, its relationship with visual attention and the image features relied upon by observers to rate it, is paramount to its successful modelling.

4.2.2 Image Features Affecting Realism

As discussed in Section 2.4, visual realism is affected by a multitude of features of the stimulus (Zhu et al., 2015). Related work discussed in Chapter 2 suggests that a) not all properties of images are perceived with the same reliability by human observers, and b) scene context seems to have a significant impact on the reliability of such perceptual tasks.

In addition to physical properties of the scene, local statistical differences introduced

to regions of a photograph of a scene can also affect realism. This is illustrated by Xue et al. (2012), who analysed the statistical relationships between image features of objects and surrounding scenes in real photographs. They found that differences in image features, such as brightness, contrast, saturation and colour temperature had a significant impact on human realism judgments. In a different study, Fan et al. (2014) conducted a survey asking observers to rank various semantic and visual attributes in a dataset of real images and image composites. Another observer group was then asked to rank the realism of these images. Spearman’s rank correlations between the resulting realism rankings and visual attributes were then calculated, highlighting features which significantly contribute to realism rankings. Their results align with the realism taxonomy proposed by Ferwerda (2003), indicating a high correlation between realism rankings and a photographic appearance. Other significant correlations included natural lighting, colours, perspective, objects and combinations thereof, indicating the importance of naturalness - the agreement between a photographic representation of a scene and its representation in an observer’s prototypical memory. This is further confirmed by the fact that “unusual”, “strange” or “mysterious” images were found to elicit lower realism ratings. This study thus illustrates the significant impact of observer experience, showing familiarity and prior visual experience correlate positively with realism judgments.

Controlled experiments assessing the impact of local transformations on attention in natural scenes have been carried out previously, notably by Einhäuser and König (2003) and Marius’t Hart et al. (2013) who assessed the impact of local contrast manipulation in photographs of natural scenes on observer attention, using methods based on signal detection theory. While not carried out in the context of image composites, these experiments still provide a baseline of human performance in detecting local luminance contrast shifts in natural images. They find that local contrast increases, particularly those associated with objects, attract overt observer attention. However, these effects are proportional to the magnitude of the contrast shift and do not occur for small shifts, highlighting the need to establish JNDs in this domain and encode into functions for automated processes, such as compositing.

4.2.3 Synthetic Image Composites

The task of image compositing is under-constrained. In order to perfectly match the appearance of a foreground object to a background scene, it would be necessary to understand and recreate the causal factors behind the appearance of this scene through an inverse rendering process (Marschner and Greenberg, 1998). With this information, one could insert any desired object into the recreated 3-D scene and perform forward rendering in order to generate the resulting composite image. While inverse rendering pipelines do exist, they currently only perform well for appropriately constrained scenarios, such as facial re-enactment (Thies et al., 2016b). They also require 3-D models of the foreground objects, often unavailable in the case of image-based compositing. Finally, application of

this process is still likely to yield perceptually low-quality results, due to the imperfections of even state-of-the-art inverse rendering pipelines.

Due to this under-constrained nature of compositing problems, much research into realistic image compositing has been carried out under constrained conditions, in order to limit the number of factors under study (see Section 2.4 for a discussion of factors impacting perception of realism). For example, Xue et al. (2012) identified image properties highly-correlated between the foregrounds and backgrounds of natural images and then applied controlled offsets of these properties in natural images in order to simulate compositing artefacts. Starting with natural images guarantees that the manually-introduced distortions are the only variable between the real image and *faux* composite, allowing for the mapping of such objective image-based differences to subjective quality, or realism scores. The authors showed a relationship between disparities in foreground-background feature distributions and observer realism ratings, however due to the small number of images used in their perceptual studies, the results are difficult to generalise to new images. A similar approach is adopted by Lalonde and Efros (2007), who develop methods for assessment and improvement of composite realism through statistical comparison of colour distributions. This is performed both against a global statistical representation of colour in natural images, but also local co-occurrence of certain colour palettes between the object and scene.

4.2.4 Difference Thresholds

Difference thresholds, or just-noticeable differences (JNDs) provide a practical framework for mapping physical stimuli to perceptually-relevant scales. This allows for any physical stimulus to be measured with respect to human perceptual sensitivity. Formally, unit JND is the amount by which a given stimulus must be changed in order for a difference to be detectable at least 50% of the time. JNDs are used extensively in various applications, including perceptual metrics (Ferzli and Karam, 2009), perceptually-based image and video processing (Jia, Lin and Kassim, 2006), colour image compression (Chou and Liu, 2008) and various models of vision (Lubin and Fibush, 1997). Accordingly, the JND can be used as a universal unit of measurement of perceptual distances.

4.2.5 Towards Generalisable Models

Existing models of visual realism suffer from similar problems to the classical IQA methods discussed in Chapter 2: they are commonly based on only a particular subset of image features or distortions, require considerable time effort and a large pool of observers, or computationally-expensive models of the HVS. Some of the most intuitively explainable models, such as that proposed by Fan et al. (2014) or Fan et al. (2018), while clearly describing the relationship between semantic properties of images and subjective visual realism, significantly under-perform relative to humans when applied to related image tasks, such as binary classification of realism, or localisation of relevant

distortions. One of the main challenges of such semantically-based subjective models is the mapping of qualitative visual features (e.g. ‘naturalness’) to image features. The error sensitivity framework, used with considerable success in IQA (Wang et al., 2004), offers an alternative solution to this problem by modelling sensitivity to different image distortions or transformations. It allows for sensitivity to any type of distortion to be modelled using a unified experimental methodology, making such an approach easily transferable and extensible to other distortion types. This also enables the problem of visual realism prediction to be broken down to component parts - the detection of relevant features, their weighting, aggregation and mapping to a perceptual scale. Finally, the inherent variability of subjective human judgements and the impact of scene content pose a significant challenge for deterministic models, due to their under-constrained nature. The use of probabilistic modelling provides an elegant approach to incorporating this implicit variability, without the requirement for explicit modelling of its many causes. This, in turn, allows for visual sensitivity baselines to be established, by aggregating over groups of observers.

4.3 Methodology

4.3.1 Overview & Motivation

The purpose of this study is twofold. Firstly, it aims to model the baseline group sensitivity of human observers to local distortions common to image composites, in the context of subjective realism judgments. Secondly, it aims to produce insight into the process undertaken by observers judging image composite quality, or realism, and determine how distortion visibility impacts subjective perception of visual realism. Prior studies of visual realism (discussed in Section 4.2) have focused primarily on global realism assessment, focusing on tasks such as classification of real and computer-generated images. In such studies, observers tend to assess the image as a whole. Instead, the work presented here focuses on assessing the realism of *combinations* of objects and scenes - a fundamental property of any image composite.

As detailed in Chapter 2, psychophysics provides practical methods for quantifying relationships between physical stimuli and the associated perceptions evoked in human participants. Specifically, the popular two-alternative forced-choice (2AFC) paradigm provides an unbiased experimental framework for the estimation of JNDs. Here, this paradigm is adopted in order to estimate average group JNDs for a set of local distortions common to image composites. This is accomplished by requesting observers to perform a binary discrimination task between a real image and a synthetic composite version of that same image, for a range of images and transformation magnitudes. Average observer performance for a set of transformation magnitudes is then fit with a sigmoidal function, and the JND is estimated as the distortion magnitude corresponding to the half-way point between guessing rate and perfect performance. A statistical evaluation of the results and

N observers	Feature	Min scale	Max scale	Min offset	Max offset
25	exposure	0.1	1.9	-	-
25	contrast	0.47	2.27	-	-
25	CCT	-	-	-200	+200

Table 4.1: The parameter ranges used to generate stimuli for the experiments. Each of the three feature ranges is linearly interpolated into 11 stimulus values.

models is then presented, and the models are applied to a range of images. This is followed by a discussion of findings.

4.3.2 Experimental Design

The experiment uses a procedure based on the 2AFC design (discussed in Chapter 2) in order to estimate generalised JNDs for three types of local, object-based image transformations, representing common feature mismatches found between the elements of image composites: exposure, contrast and CCT. These JNDs and the psychometric functions they are extracted from, describe average observer discrimination performance as a function of relative feature offset magnitude, across a range of images depicting indoor scenes. For each of the 3 stimuli types, observer performance is measured for a set of 11 feature offset magnitudes, across 165 base images. The details of stimulus ranges can be found in Table 4.1.

4.3.3 Synthetic Composite Generation

In order to generate synthetic composites \tilde{I} , we follow the approach of Xue et al. (2012), using a natural image I and a corresponding binary mask M :

$$\tilde{I} = I \odot (\max(M) - M) + f(I, x) \odot M \quad (4.1)$$

where $f(I, \theta)$ is a global transformation on image I , parameterised by θ and \odot refers to the Hadamard product (Horn, 1990). This allows for a range of transformations to be applied to the object region, without affecting the rest of the scene. Figure 4.1 illustrates the process and component elements of a synthetic composite.

4.3.4 Composite Feature Selection

During generation of synthetic composites, three image features representing common local composite distortions are selected for the transformation f in Equation 4.1: *exposure*, *contrast* and *correlated colour temperature*. Each of these features represents a commonly-occurring feature mismatch resulting from different conditions under which the foreground and background of a given composite were captured. Exposure and contrast approximate the effects of varying illumination conditions, while correlated colour temperature represents colour variability due to illumination chromaticity differences

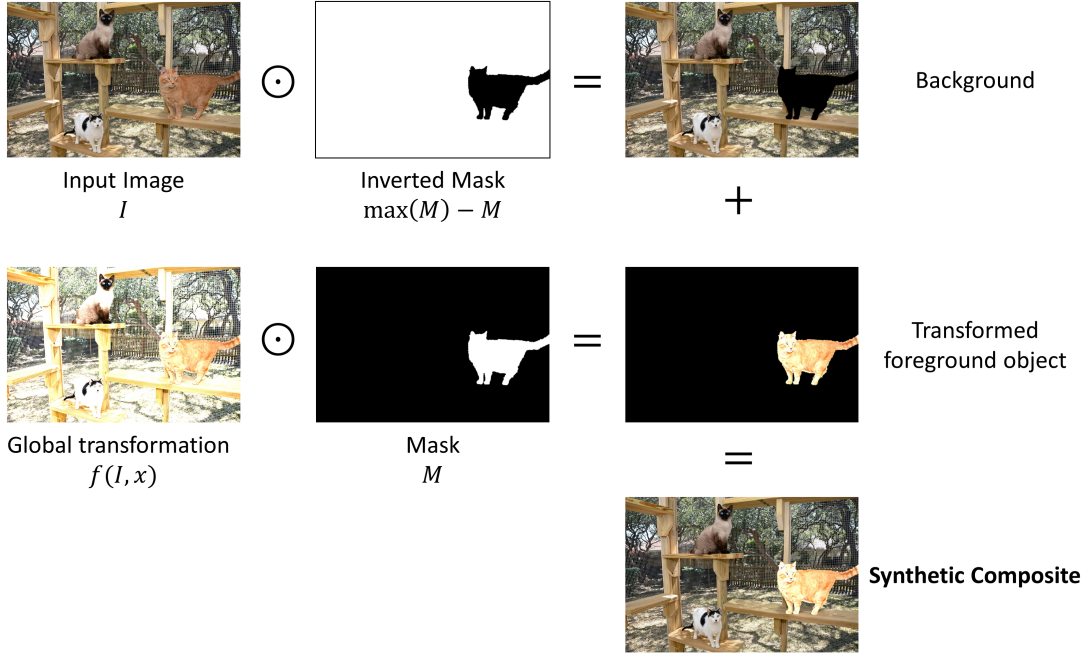


Figure 4.1: Illustration of the synthetic composite generation process. A global transformation $f(I, x)$ is first applied to the input image. The mask M is used to separate the foreground and background elements, which are then combined through pixel-wise addition. \odot refers to the Hadamard product (Horn, 1990)

between the composite elements. Both these properties (natural colour appearance and natural illumination) are highly correlated with realism, based on Fan et al. (2014, 2018). Previous work by Xue et al. (2012) has also identified these features as being significantly correlated between the foregrounds and backgrounds of natural images. Local transformations of these features are formalised below, assuming 8-bit images with intensities in the range of 0 – 255.

Exposure

In terms of image acquisition, exposure is the amount of light per unit area of the image sensor, measured as a product of local illuminance and exposure time. Exposure thus encodes both scene illumination incident on the sensor, and some settings of the camera. According to Wright (2013b), in manual compositing, correction of the distribution of pixel intensities is often the first and most important step to achieving a high quality compositing result. Adjustment of exposure is one such operation, affecting both apparent brightness, and contrast.

In this study, exposure shifts are implemented by scaling of pixel intensities in the V channel of the image converted to HSV colour space. Only pixels belonging to the foreground element are affected, as indicated by the binary mask M :

$$\hat{V}_{i,j} = \begin{cases} V_{i,j}, & \text{if } M_{i,j} = 0 \\ aV_{i,j}, & \text{otherwise} \end{cases} \quad (4.2)$$

Here a is the value of the exposure scalar, $V_{i,j}$ is the pixel intensity at row i and column j of the V channel of the input image, $M_{i,j}$ is the value at row i and column j of a binary mask, where the truth value indicates pixels belonging to the foreground object. $\hat{V}_{i,j}$ is the new pixel intensity at row i and column j after application of the transformation.

In this study, exposure shifts are expressed in *stops*, where 1 *stop* is equivalent to doubling the intensity of a pixel:

$$stop = \log_2 \left(\frac{2 \times V_{i,j}}{V_{i,j}} \right) \quad (4.3)$$

Figure 4.2a illustrates the effect of applying local exposure shifts to an example image.

Contrast

Contrast describes the difference between the highest and lowest intensities in a given visual stimulus. In the context of composite distortions, contrast can simulate disparities in dynamic range and camera post-processing. In this study, contrast shifts correspond to scaling pixel intensities in the V channel of an image in HVS colour space. Unlike exposure, intensities are scaled around the middle grey value $g = 128$. Only pixels corresponding to the foreground element are affected, as indicated by the binary mask M :

$$\hat{V}_{i,j} = \begin{cases} V_{i,j}, & \text{if } M_{i,j} = 0 \\ b(V_{i,j} - g) + g, & \text{otherwise} \end{cases} \quad (4.4)$$

Here b is the value of the contrast scalar, $V_{i,j}$ is the pixel intensity at row i and column j of the V channel of the input image, $M_{i,j}$ is the value at row i and column j of a binary mask, where the truth value indicates pixels belonging to the foreground object.

Similarly to exposure shifts, the contrast of an image can be expressed in stops, based on its contrast ratio:

$$DR = \log_2 \left(\frac{V_{\max}}{V_{\min}} \right) \quad (4.5)$$

where V_{\min} and V_{\max} are the minimum and maximum intensity of the V channel of an image in HVS colour space.

In the case of a conventional 8-bit digital image, with a contrast ratio of 256 : 1 the maximum DR is equal to 8 stops:

$$DR(256 : 1) = \log_2(2^8) = 8 \text{ stops} \quad (4.6)$$

Therefore, scaling the intensities of an 8-bit image by 50% around the mean intensity will reduce the dynamic range by 2 stops, while scaling it by 200% would increase it by 2 stops, however, this would result in clipping.

Figure 4.2a illustrates the effect of applying local contrast scaling to an example image.

Correlated Colour Temperature

Correlated colour temperature (CCT) is a measure of illuminant chromaticity and is related to the spectrum of light emitted by a theoretical blackbody at a particular temperature, measured in degrees Kelvin (Borbély, Sámson and Schanda, 2001). Illuminants with lower temperatures emit reddish light, moving through orange, yellow, white to blue, as the temperature increases. For example, candlelight (1500K) has an orange-yellow colour cast, compared to an overcast sky (6500K), which in turn appears more blue. As not all light sources are perfect theoretical blackbodies, or even incandescent lights, *correlated* colour temperature is used as a way of describing the temperature a theoretical blackbody radiator would need to reach to emit a given colour of light. While the HVS accommodates to changing illumination chromaticities, in photography, corrections must be made to avoid excessive colour casts. This is accomplished by setting an appropriate white balance, given the average CCT of the illuminants in the scene, resulting in a neutral appearance of white in the image. While CCT can be expressed in Kelvin, a perceptually-aligned unit of CCT measurement is *mired*:

$$m = \frac{10^6}{T} \quad (4.7)$$

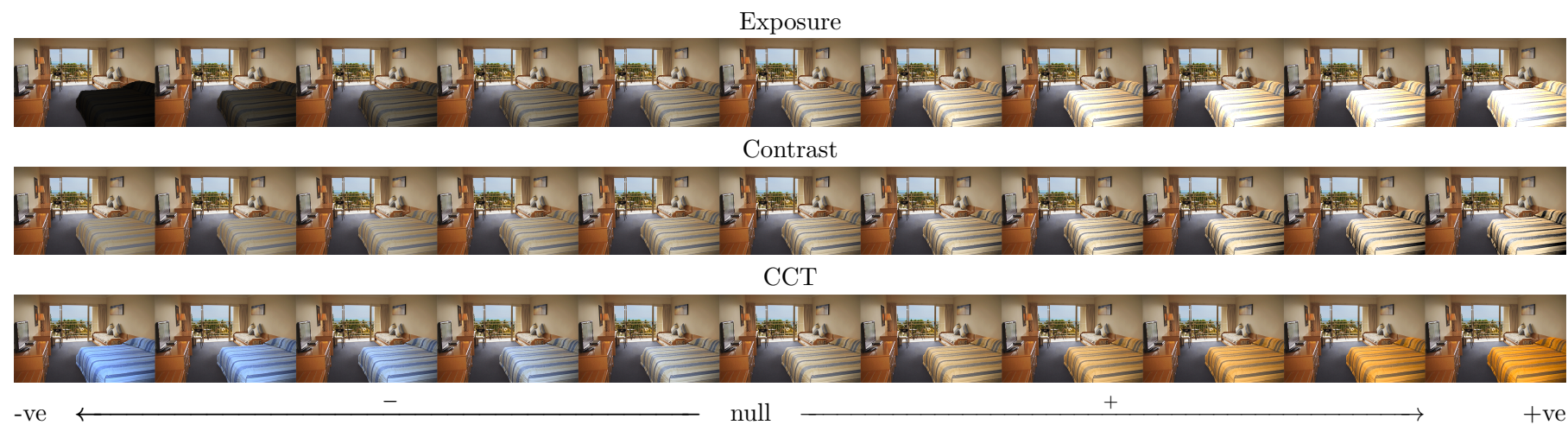
where m is the mired value, T is the colour temperature in Kelvin. The Robertson method (Robertson, 1968) is used to convert from CIELUV colour space to CTY (CCT, tint and luminance). The conversion process is a table lookup and interpolation on the Planckian locus (Wyszecki and Stiles, 1982) using the implementation by Wagberg (2020). CCT offsets are then applied by linearly shifting the mired values for each pixel. Only pixels corresponding to the foreground element are affected, as indicated by the binary mask M :

$$\hat{C}_{i,j} = \begin{cases} C_{i,j}, & \text{if } M_{i,j} = 0 \\ C_{i,j} + m, & \text{otherwise} \end{cases} \quad (4.8)$$

where m is the amount of shift in mired, $C_{i,j}$ is the CCT value at row i and column j of

the *CCT* channel of the input image converted from RGB using the Robertson method (Robertson, 1968).

CCT is a convenient method to statistically describe the colour of illumination in a scene and simulate disparities in scene illuminant chromaticity. Figure 4.2a illustrates the effect of applying local CCT shifts to an example image.



(a) An example of the stimuli used in the experiments. Top row: exposure scaling from 0.1 to 1.9; Middle row: contrast scaling from 0.43 to 2.27; Bottom row: CCT offsets from -200 to +200 mired).



(b) The object mask used to generate the images in Fig. 4.2a

Figure 4.2

4.3.5 Base Image Dataset

For the purpose of experimental stimuli generation using the features described in Section 4.3.4, a dataset consisting of 165 manually-segmented, natural images is sourced from the SUN Database (Xiao et al., 2010). These images are selected manually by an expert compositor to represent a wide range of everyday scenes, in order to cover a range of indoor scenes and objects representing natural, everyday circumstances. The segmented objects were selected to cover a wide range of relative object sizes. Moreover, manual selection allows for elimination of unnatural or visually corrupted images. Figure 4.3 shows some examples of selected images and corresponding binary masks indicating pixels belonging to the segmented objects.

4.3.6 Dataset Statistics

In addition to manual filtering, a statistical analysis of the dataset is performed, illustrating the distributions of high-level image properties across the entire set of images and objects within them. As indoor scenes can be affected by artificial illumination, care must be taken to ensure that bias is not introduced, for example, through oversampling scenes with low illumination, or a particular illuminant chromaticity. This is particularly important in the experiments discussed here, since both the colour and the intensity of the apparent illumination are modified in the synthetic composite generation process.

Figure 4.4 illustrates some high-level statistical colour properties of this dataset, namely the mean RGB values of each image in the dataset (Fig. 4.4a) and their projections into two different colour spaces: HSV (Fig. 4.4b) and Lab (Fig. 4.4c). In each of these figures, the colour of each point shows a visual representation of the mean RGB values of each image, while its coordinates illustrate its values within the new colour space. The impact of scene content and illumination chromaticity is visible here, with most mean colours ranging between light blue, grey and light orange, tracing common illuminant colours. Figure 4.5 shows the images corresponding to the orange (Fig. 4.5a), grey (Fig. 4.5b) and blue (Fig. 4.5c) X in each of the scatter plots. These images illustrate that even images lying far from the centre of the dataset (in terms of mean RGB values) represent everyday indoor scenes. Furthermore, a Shapiro-Wilk test performed on each of the distributions of mean L , a and b values across the dataset fails to reject the hypothesis that samples come from a normal distribution. This suggests that the dataset are not unrepresentative of natural scenes, and covering a range of average illuminant colours and intensities.



Figure 4.3: Examples of images used in the experiments. The binary masks in the top-left corner of each image show the object chosen for processing.

Object Size Distribution

The segmented objects vary in size, measured as a function of total image area. Since the effective size of visual patterns has a direct impact on perception of details and consequently distortions, the relative size of the object must be sufficient for an observer to extract relevant information from. Equally, the object should not occupy the majority of the scene, to allow for contextual appearance information to be assessed by observers. As this study focuses on the relationship between the appearances of the object and surrounding scene, the majority of the foreground objects in the dataset occupy less than 50% of the total image area (see Fig.4.6).

Dataset Limitations

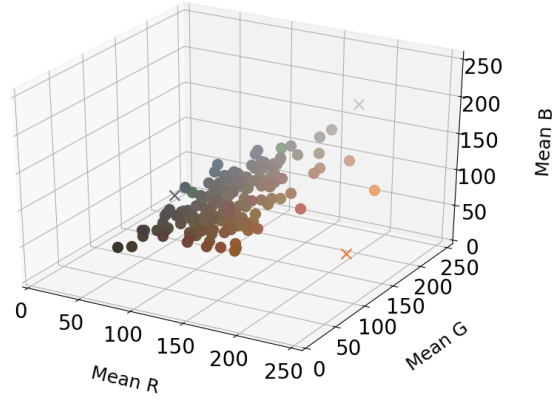
While the dataset presented here is collected to best represent transformations common to image composites, it does not represent all possible transformations. For example, while authentic image composites may include distortions of the pose, orientation, or support of composited objects, the transformations introduced in this dataset only include changes in the luminance and colour properties of the segmented objects, while not considering transformations such as changing the position, or orientation of a particular object. Accordingly, any models developed using this dataset will only be relevant to such transformations. Furthermore, the distribution of object classes present in the dataset is constrained by the sampling strategy adopted by the authors of the original dataset.

4.3.7 Apparatus & Task

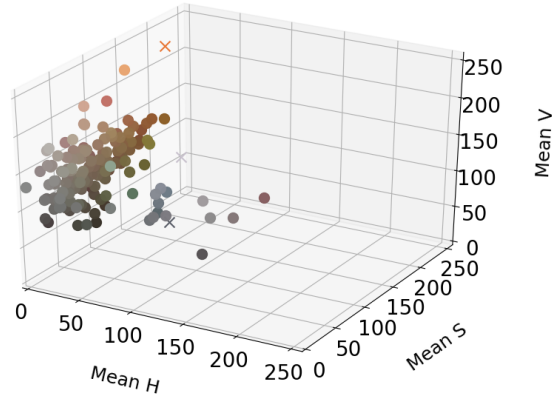
The experimental environment is designed according to viewing conditions for subjective assessments in laboratory conditions, recommended in Section 2 of the ITU Recommendation BT-500 (ITU, 2002). This includes instructions for calibration of display characteristics, ambient illumination, as well as specifications for experimental apparatus and presentation of stimuli. During the experiment, each observer is required to view 165 2AFC stimulus presentations, selecting the most *realistic* of the two images presented, based on the appearance of the object indicated by the binary mask.

Display & Environment

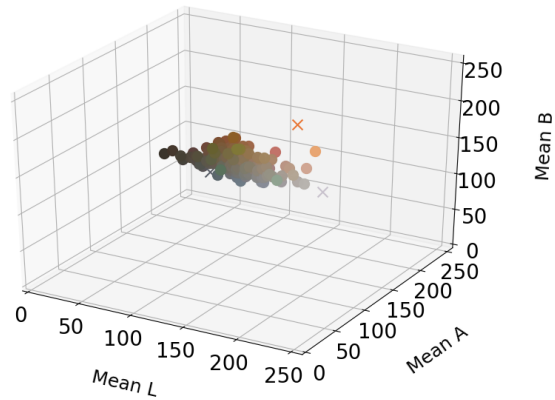
A self-calibrating Eizo ColorEdge CG247 monitor with a resolution of $1920 \times 1080px$ was used. The display was set to the factory-calibrated sRGB profile and positioned in a darkened room with no external ambient illumination. Observers were sat directly in front of the monitor at a distance of 50cm and used a keyboard and mouse during the experiment (see Fig.4.7).



(a) Scatter plot of image-wise mean RGB values.



(b) Scatter plot of image-wise mean RGB values projected into HSV colour space. The 3D position represents the coordinates of each point's RGB colour in HSV colour space.



(c) Scatter plot of image-wise mean RGB values projected into Lab colour space. The 3D position represents the coordinates of each point's RGB colour in Lab colour space.

Figure 4.4: High-level visualisation of the image dataset used in this study.



Figure 4.5: Illustration of some outliers from the image dataset. These images correspond to the a) orange, b) grey, and c) blue X markers in Figure 4.4. The image in a) represents a very orange scene, b) a bluish, mostly achromatic scene, while c) represents a dark blue scene.

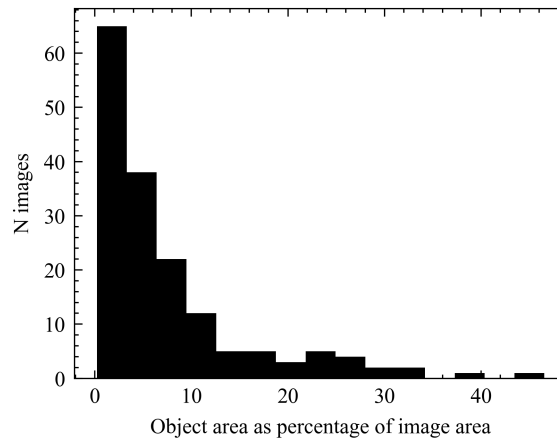


Figure 4.6: Histogram of object areas as a percentage of the entire image area for each image in the dataset.

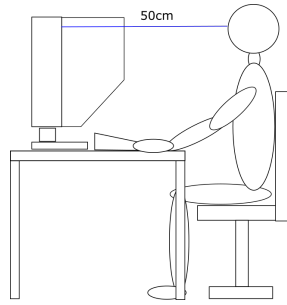


Figure 4.7: An illustration of the experimental setup. Observers use the mouse and/or keyboard to provide responses.

Hello and thank you for taking the time to complete my test. Before you continue, please take a few minutes to read this and fill in the information on the next page. Please rest assured that any personal information you provide will be kept private and will be not linked to the results of the test. Please read this page carefully, as it will detail the task you are about to undertake. If you have any questions at the end, feel free to ask the experimenter.

The experiment will run between 30 and 40 minutes (depending on how long you take to answer). Please devote your full attention and try to make sure that your answers reflect your opinion at the time (i.e. don't just vote blindly / keep pressing one button, regardless of what you think of the images).

Instructions:

- You will be presented with a set of 165 image pairs.
- Each pair will consist of two instances of the same image (usually containing a scene with some objects).
- The properties of an object in one of the two images will be altered.
- The location of the object in question will be indicated on a third image (in black & white, between the two main images).
- You will have 10 seconds to scrutinise the image pair (you will not be allowed to vote during this time)
- You will then have 5 seconds to cast your vote. You must cast a vote within this time limit.
- For each pair, you will be required to select the image that looks most realistic to you.
- To make a selection you can either click on the A or B buttons above the images OR press A or L on the keyboard.
- The differences between the images will range in intensity, and thus their ease of detection. This is not a fault with the test.
- All image pairs require a vote, before the test finishes.

Listing 4.1: The instructions presented to each observer in the experiment.

Stimuli

During each trial, observers are presented with 3 images - the real image, a synthetic composite version of that image and a binary mask indicating the pixels belonging to

the foreground object. The order of the two images is randomised every trial, and the binary mask stays in the centre (see Fig.4.8). Indicating the foreground object location mitigates observer lapses due to performing visual object search, or completing the task based on the wrong object. Furthermore, the order in which different images are displayed is randomised for each observer to mitigate learning effects. No observer sees the same image twice.

In alignment with ITU (2002), the images were displayed on an sRGB middle grey background and observers were given 10 seconds to view each image pair, followed by 5 seconds to cast their vote. The task was to indicate which of the two colour images looked more realistic. The verbatim instructions can be seen in Listing 4.3.7.

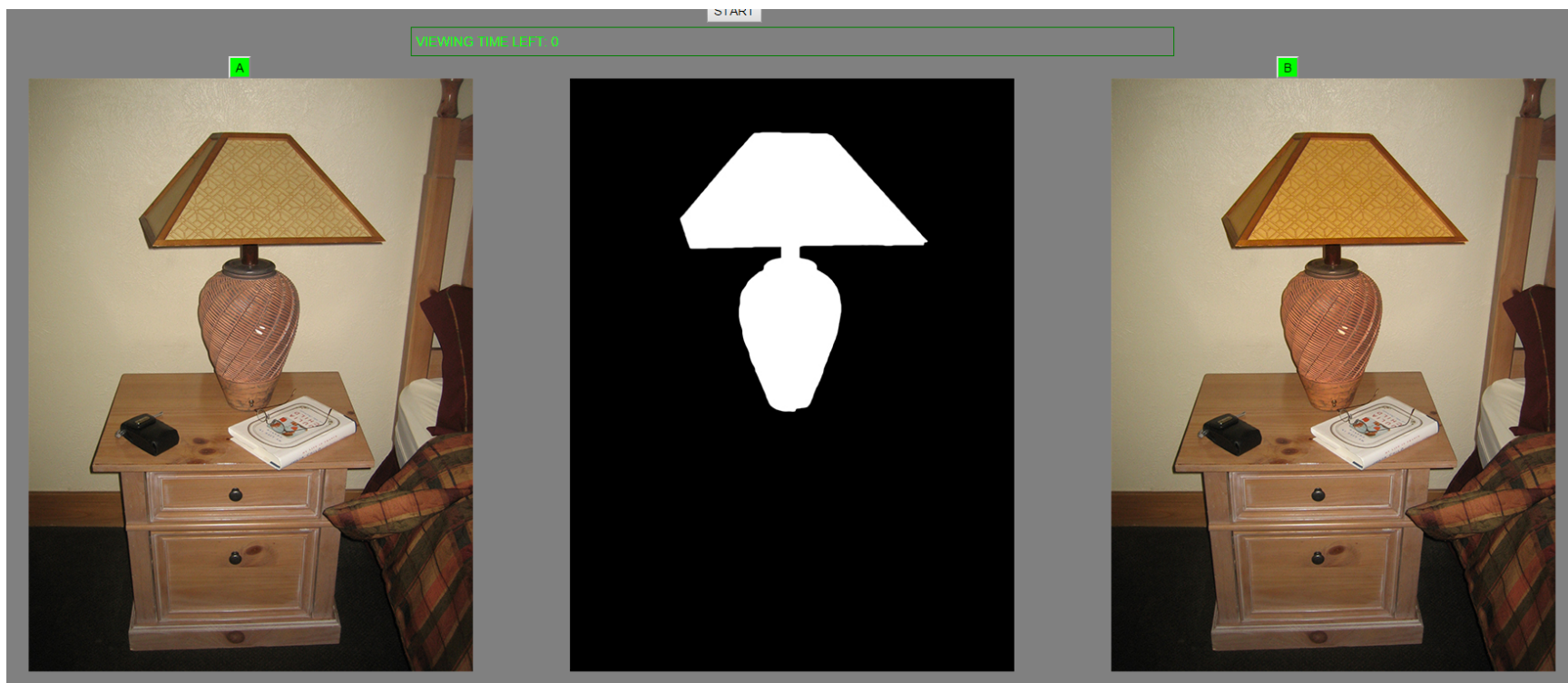


Figure 4.8: An illustration of the stimulus displayed to observers during each trial. *Left*: the original image, *Centre*: binary mask indicating the foreground object, *Right*: the synthetic composite with a CCT shift applied to the foreground object.

4.3.8 Observers

A total of 75 observers, 33 female, mean age of 28.53 ($SD = 10.54$), are recruited from a population of university staff and students. All observers are volunteers and are not rewarded in any way. Observers are then evenly and randomly distributed into three groups, one for each of the stimulus types. The selection process ensures that each observer is presented with each of the 11 stimulus levels (see Fig. 4.2a) an equal number of times, but distributed across a variety of scenes. The experiment lasts ~40 minutes per observer.

4.3.9 Analysis of Results

The analysis of experimental results follows the recommendations of ITU Report BT. 1082-1 Assembly (1990) and the procedures detailed by Wichmann & Hill (2001) Wichmann and Hill (2001a,c). First, proportions of correct responses per stimulus value are calculated. Here, “correct” is defined as selecting the original image, as opposed to the processed image. Psychometric functions are fit to the resulting data points using the *Psignifit* toolbox version 3.0 for Python Fründ, Haenel and Wichmann (2011), which implements the maximum-likelihood method presented in Wichmann and Hill (2001a). The psychometric function $\psi(x)$ describes the relationship between the probability of a correct response p , and a given stimulus intensity x . This is commonly denoted as in Equation 4.9:

$$\psi(x; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)F(x; \alpha, \beta) \quad (4.9)$$

Here, $F(x; \alpha, \beta)$ is a sigmoidal function. In this study, multiple such functions are evaluated based on their goodness-of-fit to the empirical data. Specifically, the *logistic function*:

$$F(x; \alpha, \beta) = \frac{1}{1 + \exp\left(-\frac{x - \alpha}{\beta}\right)} \quad (4.10)$$

the *Weibull cumulative distribution function*:

$$F(x; \alpha, \beta) = 1 - \exp\left(-\left(\frac{x}{\alpha}\right)^\beta\right) \quad (4.11)$$

and the *Gaussian cumulative distribution function*:

$$F(x; \alpha, \beta) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x - \alpha}{\beta\sqrt{2}}\right) \right] \quad (4.12)$$

The parameters $\alpha, \beta, \gamma, \lambda$ of ψ define the shape of the curve, and correspond respectively to its *threshold*, *slope*, lower (*guess rate*) and upper (*lapse rate*) asymptotes. The threshold α of this psychometric function describes its displacement along the abscissa. Specifically, it marks the stimulus intensity, for which the probability of a correct response is the same,

as that of a guess. Assuming that $\gamma = 0.5$ and $\lambda = 0$, α corresponds to the stimulus value yielding a .75 proportion of correct responses. The slope β describes the width, or difference, between the 95th and 5th percentile point of the sigmoid $F(x)$. In the 2AFC scenario, the guess rate γ is fixed to 0.5, as the probability of a correct guess in an n -alternative setting is $1/n$. λ represents the probability of a stimulus-independent lapse - an incorrect response, despite an arbitrarily high stimulus intensity. This value is used to scale the threshold value according to asymptotic performance. The fitting process is carried out using the *Pool-then-fit* method, adopted from Wallis et al. (2013). In order to estimate goodness-of-fit for each model, the *coefficient of determination* (R^2 measure) and deviance for each model are calculated and compared. R^2 describes the proportion of total variance in the observed data explained by the model. If \bar{y} is the mean of the observed data

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (4.13)$$

SS_{tot} is the total sum of squares:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad (4.14)$$

and SS_{reg} is the explained sum of squares:

$$SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2 \quad (4.15)$$

where \hat{y} is the prediction of the model and \bar{y} , then

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \quad (4.16)$$

Given a fit model M and a saturated model S the deviance of M is defined as

$$D = -2 \log \left(\frac{L_M}{L_S} \right) = -2(\log(L_M) - \log(L_S)) \quad (4.17)$$

where L_M is the maximum likelihood of the fit model and L_S is the maximum likelihood of the saturated model.

The model which achieves the lowest average deviance and highest average R^2 across all experiments is then selected for further analysis. For each parameter of the fit psychometric function 95% confidence intervals (CIs) are calculated using the bias-corrected and accelerated (BC_a) bootstrap method (Efron and Tibshirani, 1986), as

suggested by Hill (2001).

4.4 Results

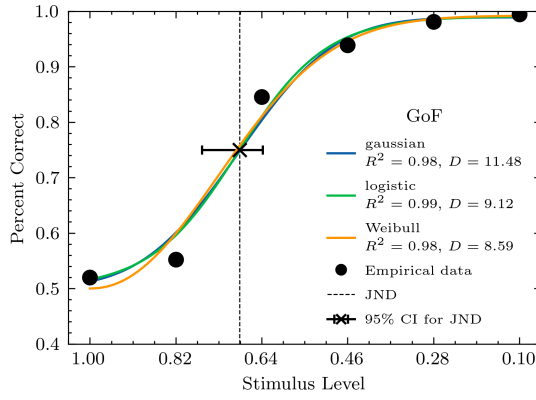
This section details the results of the experiments. A total of $165 \times 25 \times 3 = 12375$ data points are collected in the entire study, 4125 for each of the three experiments. To ensure that responses for any given image contribute equally to the model, two responses for each image-offset combination are selected, resulting in $165 \times 11 \times 2 = 3630$ total data points, 330 for each of the 11 stimulus intensities to which the perceptual functions are fit.

4.4.1 Goodness-of-fit Evaluation

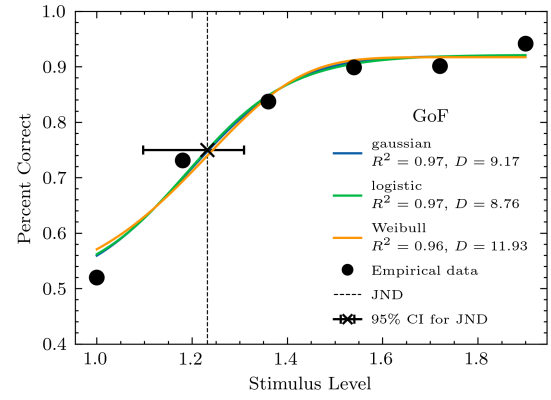
The goodness-of-fit evaluation shows that experimental data is fit well by each of the evaluated functions, achieving high R^2 scores and comparable deviance between models. Figure 4.9 illustrates each of the functions fit to mean correct response rates for each stimulus level, across each feature under test. R^2 scores and deviance measures are indicated in each legend. No significant differences between the deviance residuals were found across models for each feature, as determined by one-way ANOVA ($p > 0.05$). This suggests that the experimental data is fit well by each of the evaluated models. The following analyses use the model based on the logistic function.

4.4.2 Psychometric Functions & JNDs

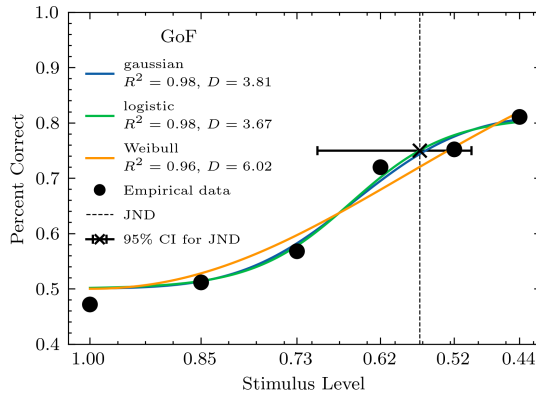
The parameters of the logistic function fit to each of the sets of experimental data can be found in Table 4.2 along with their 95% confidence intervals. The threshold values represent the average group JNDs for each feature under test. For exposure shifts, these JNDs can be expressed in stops: 0.30 and -0.54 stops for positive and negative exposure shifts, respectively. This is also true of contrast, where these correspond to 0.50 stops for positive shifts and 1.22 stops for negative shifts. JNDs for CCT are 82 mired for positive CCT shifts and -94 mired for negative CCT shifts. Lapse rates are overall highest for negative contrast offsets (19%), and lowest for negative exposure offsets (1%). Figure 4.10 illustrates the 95% CIs for lapse rate estimates, which are widest for contrast transformations, particularly negative ones. This suggests that observers performed relatively poorly in the discrimination task for this transformation type, resulting in relatively high lapse rates and wide CIs for the JND estimate.



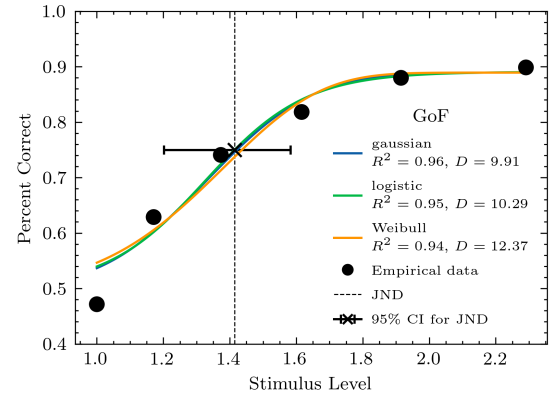
(a) Exposure (negative offsets)



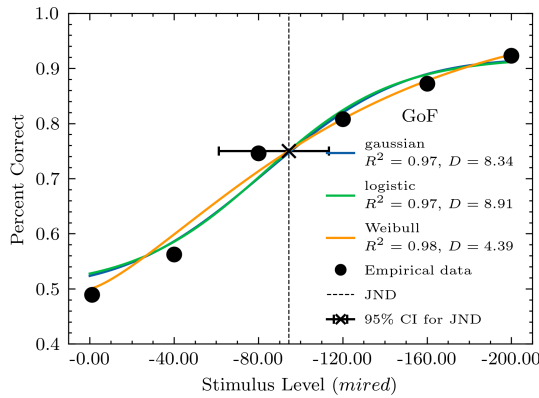
(b) Exposure (positive offsets)



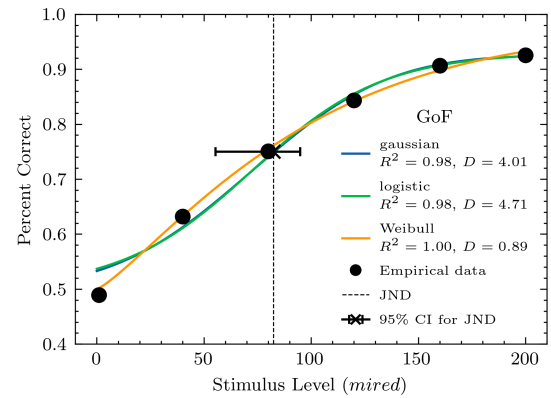
(c) Contrast (negative offsets)



(d) Contrast (positive offsets)



(e) CCT (negative offsets)



(f) CCT (positive offsets)

Figure 4.9: Goodness of fit evaluation: Each figure shows proportions of correct responses for each stimulus level (black dots). The lines indicate best-fitting **cumulative Gaussian**, **logistic** and **Weibull** functions. Negative feature offsets are shown in the left column, positive in the right. The horizontal error bars indicate 95% CIs for the stimulus level corresponding to 1 JND

Feature	$\alpha-$ (JND)	$\beta-$	$\lambda-$	$\alpha+$ (JND)	$\beta+$	$\lambda+$
Exposure	0.69 [0.64, 0.77]	0.54 [0.33, 0.94]	0.01 [0.00, 0.04]	1.23 [1.10, 1.30]	0.64 [0.38, 1.62]	0.08 [0.01, 0.11]
Contrast	0.43 [0.30, 0.50]	0.38 [0.22, 1.03]	0.19 [0.01, 0.23]	1.41 [1.20, 1.58]	0.88 [0.50, 2.43]	0.11 [0.02, 0.16]
CCT	-94.34 [61.14, 113.39]	182.92 [106.00, 349.03]	0.08 [0.01, 0.13]	82.40 [55.38, 94.86]	180.17 [120.97, 317.95]	0.07 [0.01, 0.11]

Table 4.2: Parameter values of logistic psychometric functions fit to the experimental data for each of the transformation features. Parameters follow naming convention from Eq. 4.9 and 4.12. Parameters followed by $-$ indicate fit for negative transformations, while parameters followed by $+$ indicate fit for positive transformations. Accordingly, α is the threshold of the respective psychometric functions, corresponding to $1JND$.

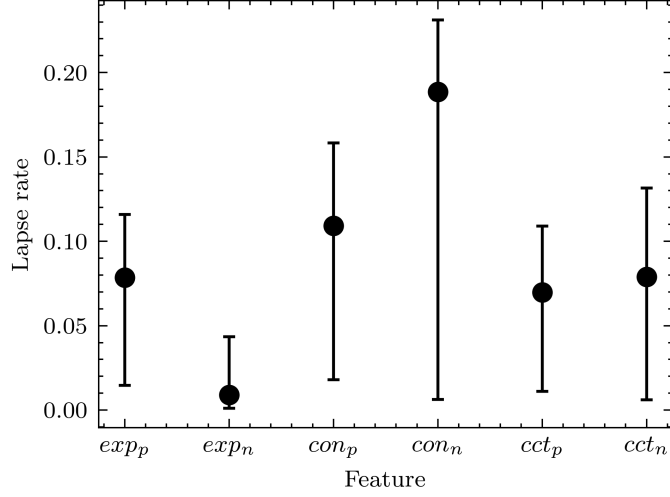


Figure 4.10: Mean lapse rates and their corresponding 95% confidence intervals for positive (subscript p) and negative (subscript n) offsets of exposure (exp), contrast (con) and CCT (cct).

4.4.3 Qualitative Evaluation

In order to illustrate the generalised JNDs, local exposure, contrast and CCT shifts are applied to a range of test images. Specifically, Equations 4.2, 4.4 and 4.8 are applied, using the JND values from Table 4.2 as respective arguments.

Finally, the average discrimination performance ADP for each image-feature combination is used to recover ‘difficult’ image-feature combinations, for which average discrimination performance was lowest, and ‘easy’ image-feature combinations, for which average discrimination performance was highest. This is performed by counting all correct responses for a particular image-feature combination and dividing by the total number of responses for that combination, as in

$$ADP = \frac{1}{n} \sum_{i=0}^n r_i \quad (4.18)$$

where n is the number of responses for a given image and transformation feature and r_i is the i th response. Correct responses are encoded as 1 while incorrect ones as 0, therefore if all responses for a given image-feature combination were correct, then $ADP = 1.0$. Visual analysis is then performed in order to illustrate common properties of these examples, which may impact observer sensitivity.

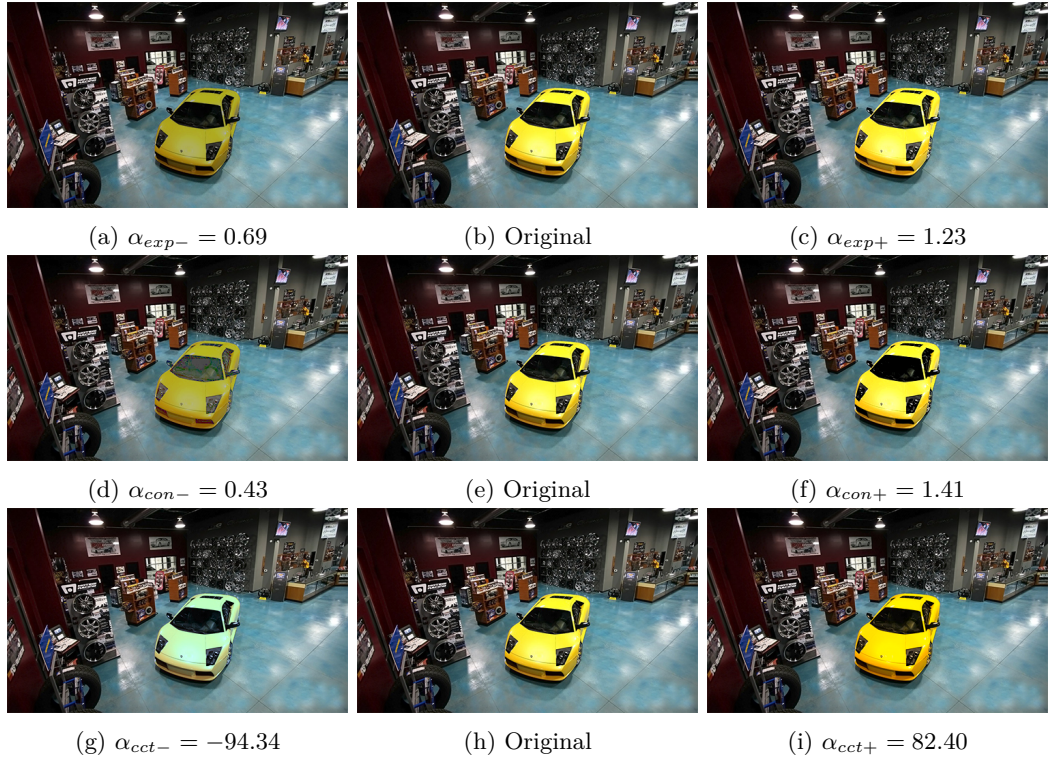


Figure 4.11: The exposure, contrast and CCT JNDs visualised for an image from the experimental dataset. The middle column contains original images, while the left and right columns contain negative and positive JNDs for exposure, contrast and CCT offsets respectively, top to bottom.

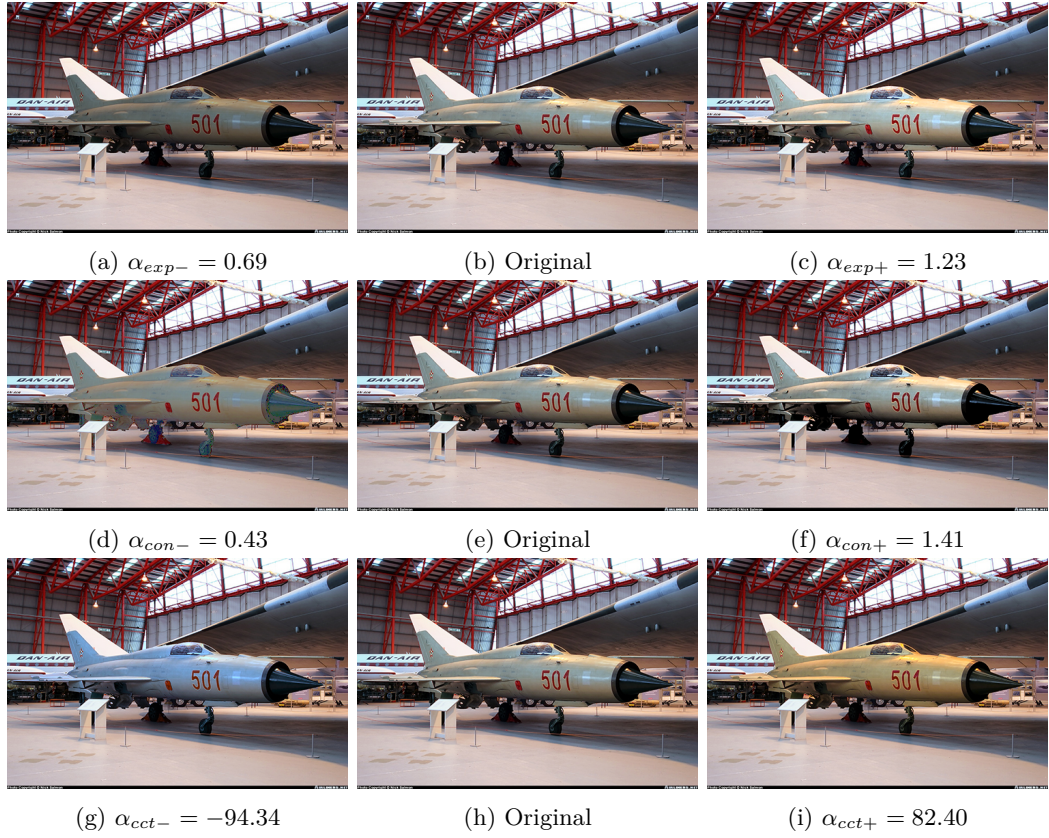


Figure 4.12: The exposure, contrast and CCT JNDs visualised for an image from the experimental dataset. The middle column contains original images, while the left and right columns contain negative and positive JNDs for exposure, contrast and CCT offsets respectively, top to bottom.

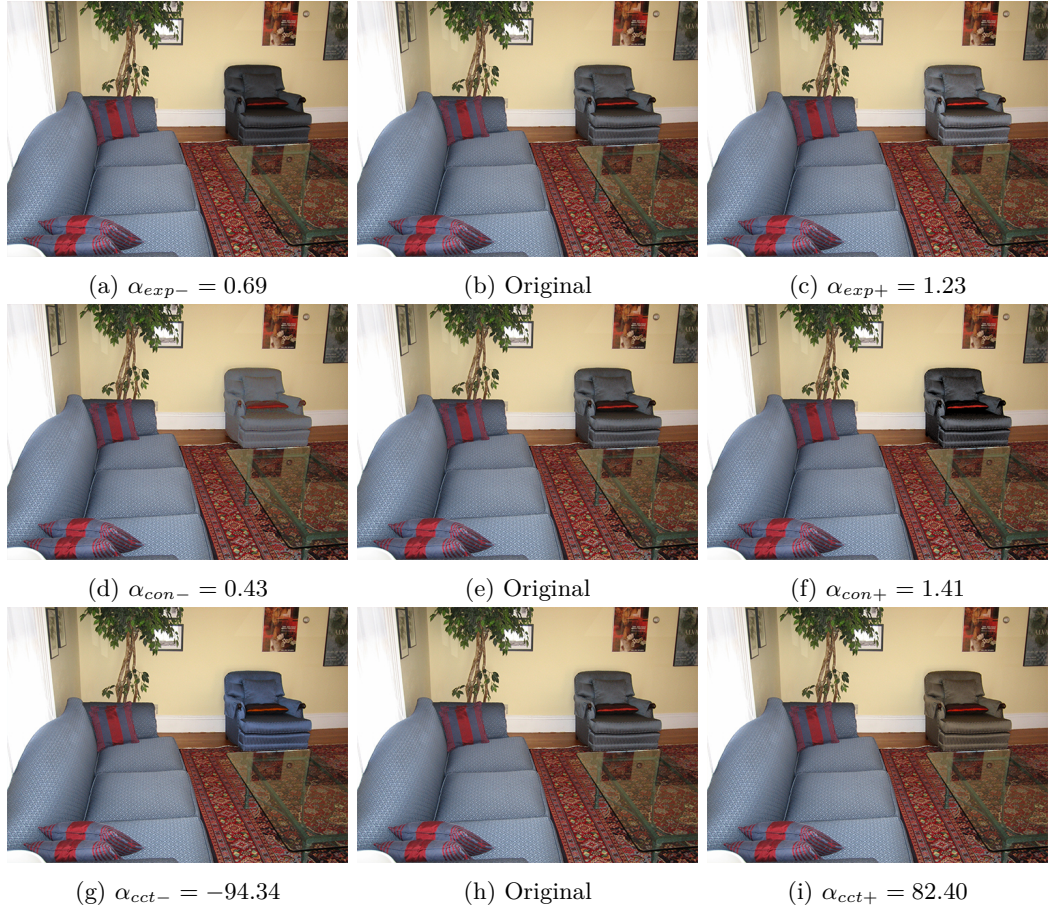


Figure 4.13: The exposure, contrast and CCT JNDs visualised for an image from the experimental dataset. The middle column contains original images, while the left and right columns contain negative and positive JNDs for exposure, contrast and CCT offsets respectively, top to bottom.

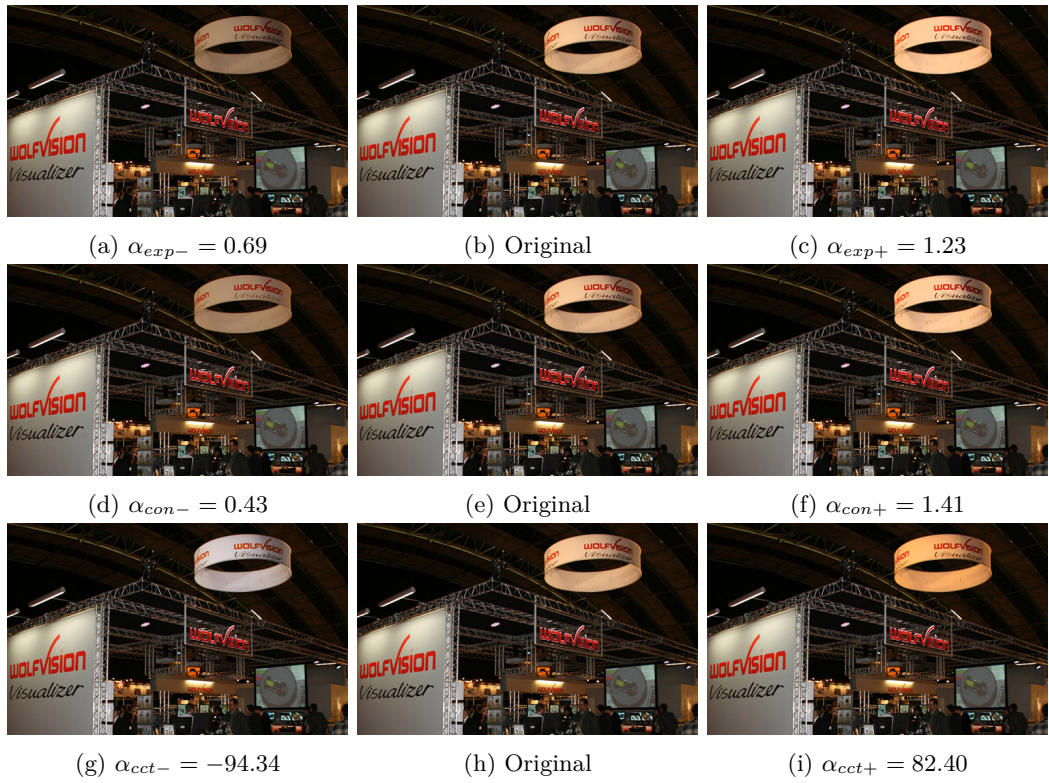


Figure 4.14: The exposure, contrast and CCT JNDs visualised for an image from the experimental dataset. The middle column contains original images, while the left and right columns contain negative and positive JNDs for exposure, contrast and CCT offsets respectively, top to bottom.

4.5 Discussion

4.5.1 Overview

The results indicate that psychometric functions based on foreground-background distortion perception paradigm can be successfully used to model perception of visual realism in image composites. The experimental data are fit well by the proposed models, and while some differences in observer performance are present between exposure, contrast and CCT, the results obtained here are in line with previous work. Similar to Xue et al. (2012), who evaluated the impact of foreground-background disparities, subjective visual realism decreases as the disparity magnitudes increase. The psychometric functions obtained can serve as probabilistic models to compare the relative change in exposure, contrast or CCT required for observer detection. JNDs obtained in these experiments also provide a perceptual scale for mapping between features and observer detection probability, and accordingly serve as a predicate of their realism rating.

4.5.2 Observer Performance & Lapse Rates

Among all features evaluated, contrast offsets, particularly negative ones, yielded the highest lapse rate estimates and widest lapse rate CIs (see Fig. 4.9c and 4.9d, and Table 4.2). Similar results have been obtained by Einhäuser and König (2003) who indicated that a 60% reduction in peak contrast resulted in $78\% \pm 6\%$ correct response rate, while an 80% increase in peak contrast resulted in an $81\% \pm 9\%$ correct response rate. The contrast model developed in this study falls within this range, yielding 72% correct response rate for the negative offset and 84% for the positive one.

The high lapse rates for negative contrast offsets also mirror the results of Marius’ Hart et al. (2013), who show that contrast in natural images is correlated with attention, and decreases in contrast, particularly those applied to object regions, reduce fixation and detection probability. The same is not true for relative increases of contrast, which is reinforced in the work presented here. Thus, when matching the contrast of an object to that of a scene, underestimating object contrast is likely to appear less unrealistic to an observer than overestimating it, which will in turn attract attention to that object. This is reflected in the lapse rates for negative contrast offsets - performance at 50% peak contrast compared to 200% peak contrast is $\sim 10\%$ lower.

Both CCT and exposure covered offset ranges adequately, receiving 100% correct responses for the highest offsets, in the case of some observers. Higher lapse rates for positive, compared to negative exposure offsets can be explained by a slight difference in stimulus ranges indicated by the pilot experiment. Additionally, it seems that some extreme CCT offsets can be interpreted as plausible differences in object reflectance, increasing the lapse rates, while still appearing realistic. The variability in responses for each offset level can be attributed to image and object changes. This is consistent with studies by Tan et al. (2015) and Xue et al. (2012), who found significant differences between the consistency

of ratings for different images, as well as across participants. Through the use of a larger dataset of 165 images, this work also indicates how much variability can be expected across general composites. This subjectivity of realism judgements is further illustrated by the threshold CIs in Table 4.2.

4.5.3 Qualitative Analysis

Both visualisation of generalised JNDs, and the analysis of *ADP* for different image-feature combinations, provide further insight into the sources of variability in the resulting models. One example of this can be seen in Figures 4.11-4.14, where average group JNDs are applied to a selection of images from the experiment. In each of these figures, each row represents a different feature (exposure, contrast and CCT, respectively), while the columns represent feature offsets based on average group JNDs (negative 1 JND, no offset, positive 1 JND). Comparing the same offsets applied to different object-scene combinations, it is clear that the offsets are easier to distinguish in some object-scene combinations, while being almost indistinguishable in others. The increase in CCT in Figure 4.14i is barely noticeable, while the same offset applied to the jet, seen in Figure 4.12i, is more pronounced. Similarly, the contrast shifts visible in Figures 4.13d and 4.13f are easier to notice than those in Figures 4.11d and 4.11f. This suggests that the visibility of such distortions is a function of both the original appearance of the object, as well as the type and magnitude of the transformation applied to it.

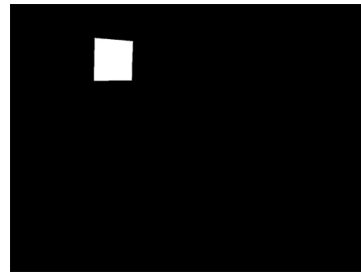
Inspection of images with highest and lowest *ADP* provides further insight into the properties of objects, which may influence detection of local transformations. Low *ADP* contributes to higher lapse rate estimates in the proposed models, but can also shed light on scene features which may be influencing this. Similarly, reviewing images for which *ADP* is consistently high, may illustrate which image features may contribute to the successful detection of such image transformations.

Object Texture

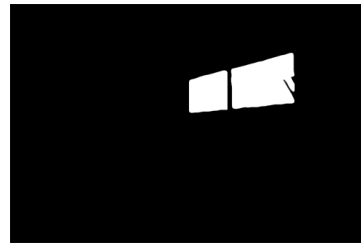
Both exposure and contrast transformations were detected less reliably when applied to uniformly-coloured and low-contrast objects. This is visible when looking at the objects in Figures 4.16 and 4.18. Each of these figures represents five scenes for which *ADP* was lowest, when the transformed feature was exposure and contrast, respectively. Luminance-based transformations applied to objects such as the statue in Fig. 4.16d, the carpet in Fig. 4.18d or the sheet in Fig. 4.18e were not reliably detected by observers. This could be related to their abstract or plain appearance and/or lack of visual detail, which makes exposure or contrast transformations result in a plausible, albeit brighter or darker appearance. Consider the vase in Fig. 4.16e, which, whether made darker or brighter, would still appear plausible, since one can imagine a vase in many colours. Conversely, the paintings in Fig. 4.15c, the car in Fig. 4.15d, or the blackboard in Fig. 4.17d are less abstract, contain more visual detail, richer textures or a wider contrast range. The *ADP*



(a)



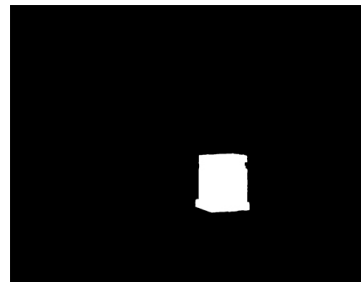
(b)



(c)



(d)



(e)

Figure 4.15: Images with the highest average discrimination performance for stimuli with exposure transformations.



Figure 4.16: Images with the lowest average discrimination performance for stimuli with exposure transformations.

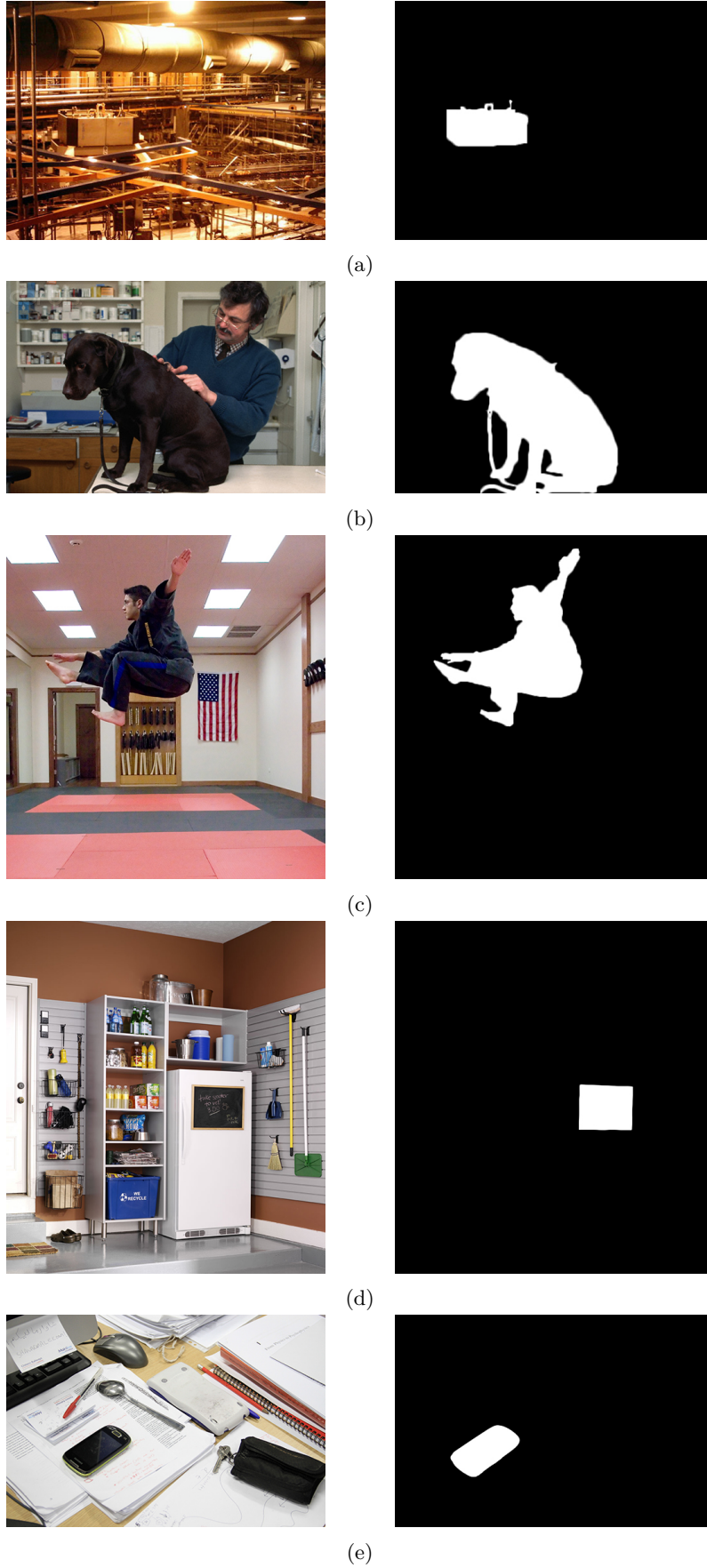


Figure 4.17: Images with the highest average discrimination performance for stimuli with contrast transformations.

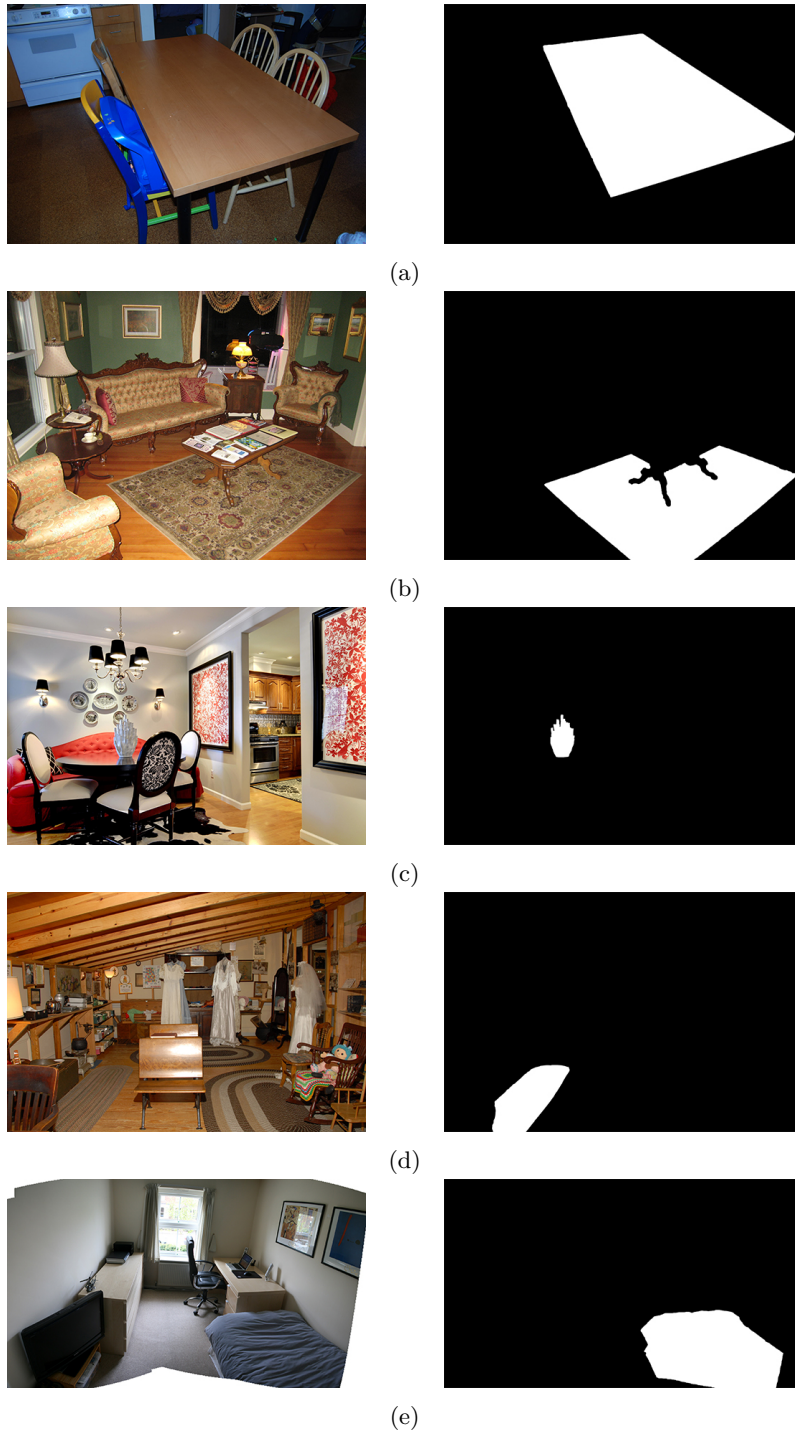


Figure 4.18: Images with the lowest average discrimination performance for stimuli with contrast transformations.

for images containing these objects is highest, in the case of exposure and contrast shifts.

Nearby Illumination

Further investigation of scenes based on *ADP* also shows that, in addition to properties of the objects themselves, sources of additional illumination near the transformed objects in the scene tend to correlate with lower *ADP*. For example, in Fig. 4.16a the laptop is directly underneath a bright desk lamp, in Fig. 4.15b the lamp is just in front of a natural source of light, in Figs. 4.18a and 4.18b the camera flash is used, while Figs. 4.18c and 4.16d contain many visible sources of artificial illumination. If, in performing the task, observers rely on internal estimation of object reflectance, then additional illumination could introduce a confounding effect, resulting in lower *ADP* for such scenes.

Contextual Inference

Scenes featuring multiple objects sharing appearance features, such as the banners in Figure 4.14, the sofas in Figure 4.13, or the container in Figure 4.17a, highlight another interesting property of this discrimination problem. Namely, observers may rely on semantically similar objects elsewhere in the scene to inform their decision. Looking at the first example, the circular banner appears most plausible when a -120 offset (bottom-left image) is applied, rendering its colour to be similar to the rectangular banner in the lower left portion of the image. Since both banners share a logo, it is plausible they should share their colour. In reality, the circular banner in the *original* version of this image must have been affected by another illumination source, compared to the rectangular one. This can be explained by the phenomenon described by Pont and te Pas (2006), whereby observers may confound the effects of illumination with object appearance. A similar scenario can be seen in the image featuring the blue sofas. Due to the variation in lighting, the background sofa appears slightly darker than the foreground one. In this scenario, an increase in exposure (top-right) may actually render the scene more plausible, due to the increased appearance similarity between the two sofas. The final example (Fig. 4.17a), featuring the industrial container, illustrates a similar scenario, with the opposite effect. Here, the uniform, orange illumination across the scene made detection of contrast transformation easy for observers, resulting in a top-5 *ADP* for this image. These examples indicate that observers may use other scene elements when inferring the true appearance of a given scene element.

Object Chromaticity

Qualitative analysis of images with respectively highest and lowest *ADP* also indicates that *achromatic* objects enabled observers to reliably detect CCT transformations applied to these objects. Figures 4.19 and 4.20 show two sets of images for which *ADP* was highest, and lowest, respectively. In the former, the majority of objects are largely achromatic, such as the plinth in Fig. 4.19b, the bear in Fig. 4.19d or the mug in Fig. 4.19e). Conversely, images with low *ADP* contain mostly monochromatic objects, commonly blue or yellow,

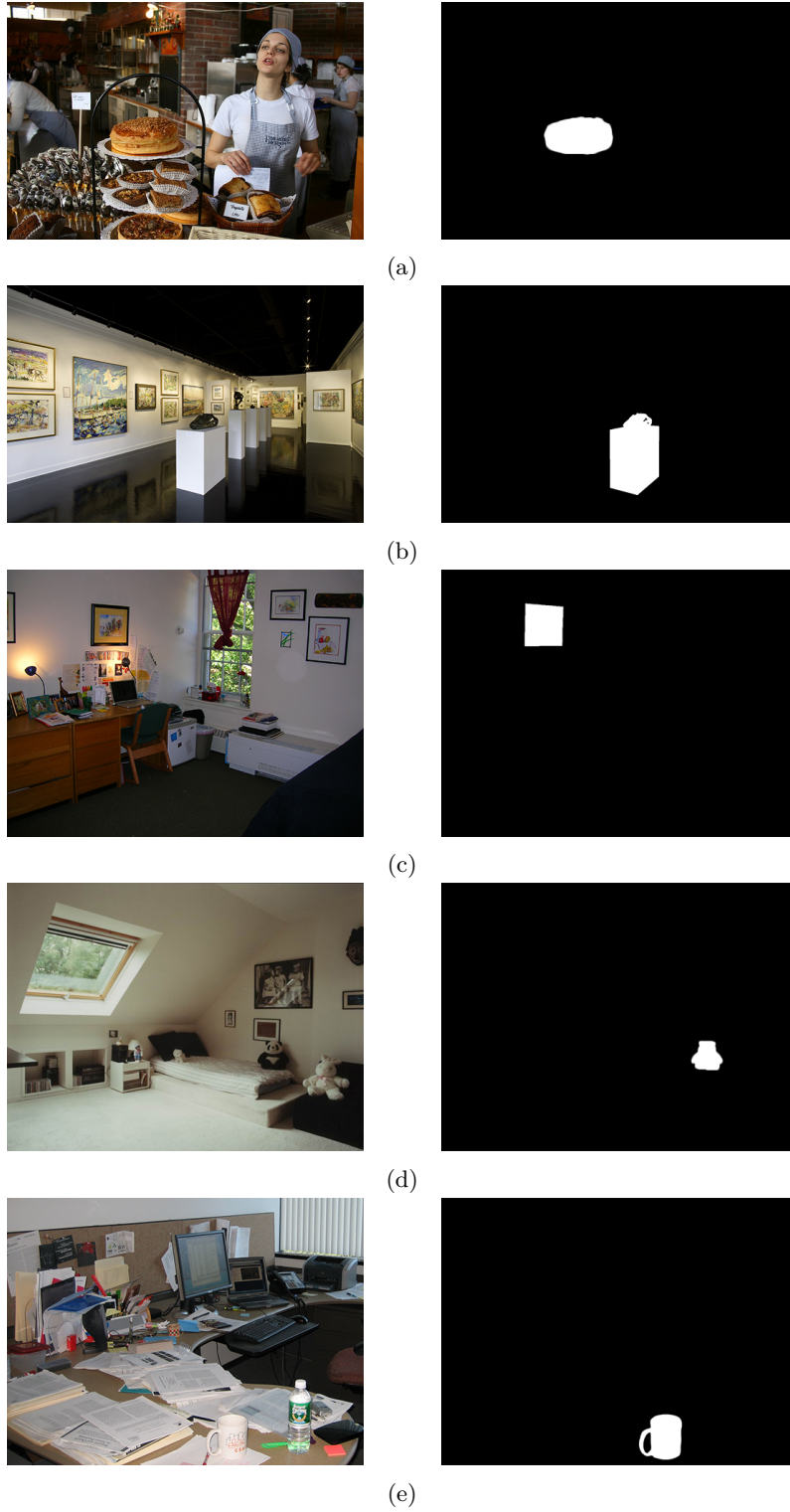


Figure 4.19: Images with the highest average discrimination performance for stimuli with CCT transformations.

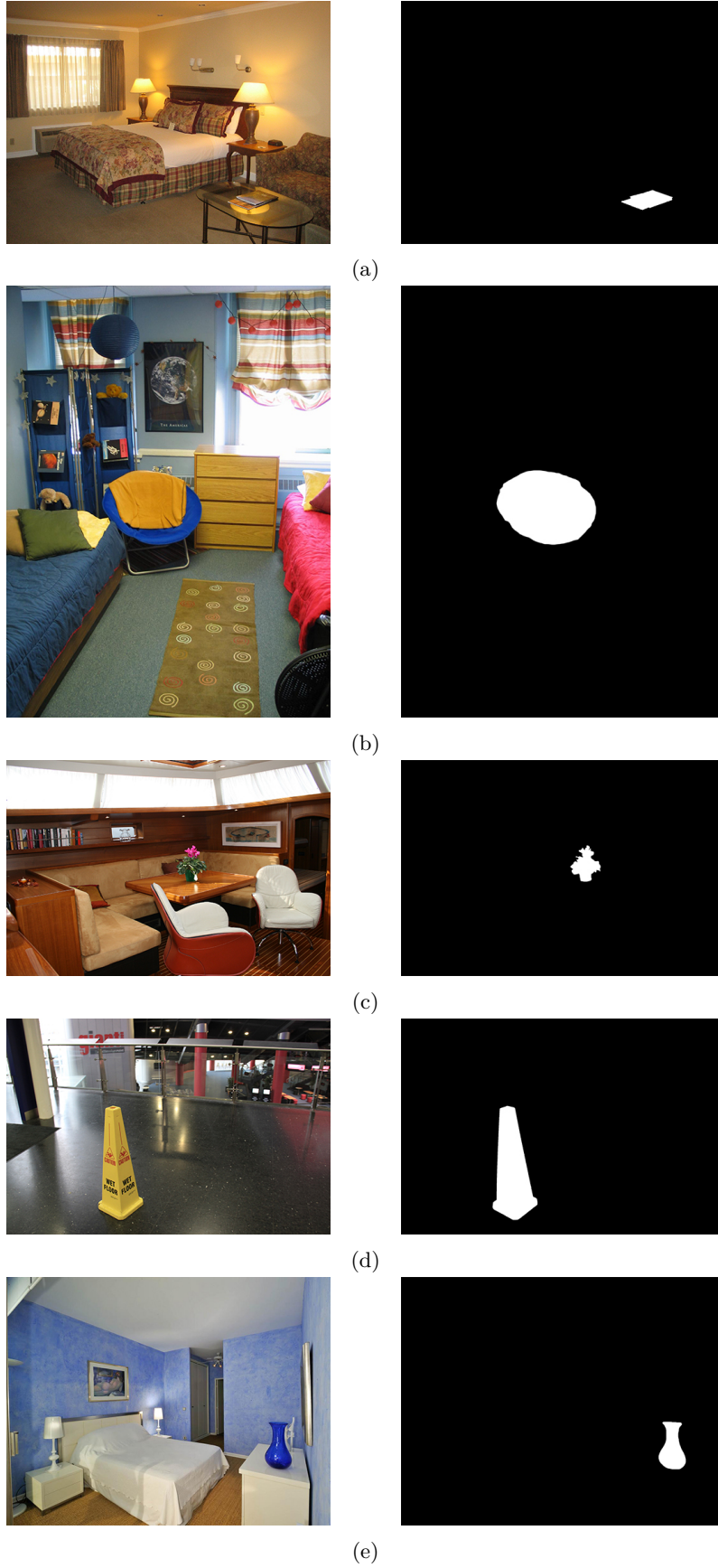


Figure 4.20: Images with the lowest average discrimination performance for stimuli with CCT transformations.

such as the seat in Fig. 4.20b, the cone in Fig. 4.20d or the vase in Fig. 4.20e.

4.5.4 Summary

Overall, the results of this study indicate that a framework based on error visibility can be successfully adopted to model visual realism in image composites. As long as a composite distortion can be approximately represented as a parametric transformation, the JND framework can be reliably used to model visual realism as a function of distortion visibility in a probabilistic setting. Importantly, the study has also gained insight into the process performed by observers when assessing the realism of object-scene combinations in natural images. The following findings have been made:

- Both the base appearance (colour, texture, illumination) and semantic properties of composite objects can interact with particular transformations, and consequently influence observer discrimination performance
- Observers may rely on different parts of the scene in order to inform their decision. This sometimes overrides the influence of sheer transformation visibility, in favour of increasing the appearance similarity of semantically similar objects in the scene
- Colour transformations applied to achromatic objects are reliably detected
- Transformations to man-made, abstract, plain or monochromatic objects may be harder to detect and appear more plausible
- Additional sources of illumination in the scene can impact the reliability of observers detecting transformations to nearby objects, particularly if these involve brightness shifts.
- Observers, properties of the object and scene, as well as transformation type can all impact resulting JNDs. As such, any predictive model for subjective visual realism should generalise across observer groups, but not over object-scene properties, instead using them as conditioning factors.

In addition, these findings also prompt several related questions:

- Are the resulting JNDs affected by explicit indication of object identity? Would natural deployment of visual attention result in different JND values?
- Do visual attention patterns differ significantly as a function of transformation type? Do observers rely on different parts of the scene to assess objects affected by different transformations, e.g. colour vs brightness?
- How can models of visual realism be conditioned on object and scene semantics?
- Which elements of the scene are relied upon to arrive at subjective realism ratings?

Based on the above, it is clear that subjective visual realism responses are not just a function of sheer impairment visibility. Rather, they are a result of a more complex process, based on inference and influenced by preference and properties of complex scenes (Yuille and Kersten, 2006; Kersten, Mamassian and Yuille, 2004). The findings presented in this chapter are further supported by the results of Tan et al. (2015), who found significant differences in observer responses for the same disparities, across different scenes and objects. Finally, it must be emphasised that subjective judgements are not always reliable. Even in an ideal scenario, where both the reference and a modified version of an image are present, and the location of the modified object is specified, observers may select the real image as the ‘unrealistic’ one, despite visible errors. For example, several observers reported difficulty in predicting the exact hue of monochromatic objects, when CCT was offset, or judging correct exposure when objects were nearby strong sources of illumination. Further research on scene perception, reflectance estimation and visual attention in natural scenes is required to understand and model these scenarios appropriately.

4.6 Conclusions

This chapter has presented a probabilistic model of visual realism, based on human perception of three key local image transformations representing common image composite distortions: exposure, contrast and CCT shifts. The findings indicate that psychometric methods and signal detection theory provide an adequate approach to modelling human sensitivity to such local distortions, and that visibility of such distortions provides a good proxy for visual realism, as indicated by the good fit of the proposed models to the empirically observed data. These findings are in-line with previous research and contribute both models and a methodology for modelling the impact of various local image transformations on subjective realism perception in image composites. The re-application of the models to the experimental dataset and investigation of average discrimination performance illustrate the considerable impact of extraneous factors, such as scene context, observer experience and attention. This shows that while relative distortion magnitude is a good predictor of visual realism, modelling the impact of scene and observer attention would likely improve the proposed models’ generalisability. Consideration should also be given to understanding the impact of providing observers with the identity of the transformed object during the experiment. It is possible that JNDs could shift if observers fail to notice the transformed object in scenarios where they are not provided with its identity.

The next chapter will further investigate this methodology in the context of visual attention and task, focusing on how local and global scene properties affect the detection and realism rating process, the role of top-down and bottom-up attention in subjective visual realism judgments, and the impact of task, prior object knowledge, and transformed image feature on gaze allocation.

Chapter 5

Impact of Attention, Task & Feature Type

This work was published in:

Dolhasz, A., Frutos-Pascual, M. and Williams, I., 2017. Composite Realism: Effects of Object Knowledge and Mismatched Feature Type on Observer Gaze and Subjective Quality. *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE, pp.9–14

5.1 Introduction

Chapter 4 presented the results of a study modelling subjective visual realism in natural image composites as a function of local transformation magnitude. Analysis of the obtained psychometric models highlighted the considerable impact of appearance properties of the object and scene, as well as the inherent variability of subjective judgments provided by human observers assessing visual realism. The design of the study also controlled for the effects of visual attention, by providing explicit information regarding the identity of target objects. In practical scenarios, however, it is often desired to consider the effects of visual attention, for example, to reduce the amount of processing applied to areas of an image composite reliably ignored by observers. Moreover, if subjective realism judgements are significantly influenced by deployment of visual attention, generalisable models of visual realism may not be achievable, without reliable modelling of visual attention. Finally, deployment of visual attention may provide indication about spatial and semantic strategies adopted by observers assessing visual realism as a function of different local transformations. This could allow for extended reasoning regarding the impact of semantic scene properties and further analysis of scenarios where transformed objects are perceived as more realistic than their original counterparts.

To investigate these issues and determine the value of visual attention modelling in composite realism assessment, this chapter adopts an eye-tracking paradigm in order to capture fixation patterns of observers judging the visual realism of synthetic composites, affected by local exposure and colour transformations. Eye tracking has been used extensively to measure overt visual attention allocation in various experimental tasks, including image quality assessment (Liu and Heynderickx, 2011). Gaze data can provide rich insight into the relative importance of different image regions to the completion of a visual task. By comparing standard gaze metrics (Bylinskii et al., 2015), such as spatial and temporal distributions of fixations, of observers performing visual realism assessment tasks, this chapter addresses the question of how visual attention, prior knowledge and transformation type affect subjective realism judgments for a range of scenes and objects. This chapter addresses the following research questions:

- Does prior knowledge of the composited object identity affect its realism judgements?
- Are different scene features attended to when the transformed feature changes?
- Do observers adopt different strategies, depending on the scene or transformed feature?

5.2 Related Work

5.2.1 Visual Attention

Typical scenes encountered by the HVS contain many objects, constituting a large amount of visual information. Due to its limited capacity for information processing, the HVS does not treat all this information with equal importance. This often results in competition between objects in the visual field for neural representation. Consequently, the HVS exhibits a property of selectivity, which enables attended information to be processed and unattended information to be largely ignored, thus conserving the limited processing capacity (Desimone and Duncan, 1995). Visual attention (VA) is the collection of mechanisms driving this selective behaviour and associated eye movements. There is also evidence indicating VA may be required to combine features of objects, and impacts short-term visual memory formation (Vecera and Rizzo, 2003). Attention has the effect of reducing observers' uncertainty in stimulus-related judgements and enhancing the perceptual representation of attended objects, compared to unattended ones. The manner in which VA is allocated between the competing stimuli depends on both the intrinsic visual properties of the stimuli, known as *bottom-up* attention, as well as the task being performed by the observer, known as *top-down* attention (Le Meur et al., 2006).

Bottom-up & Top-down Attention

Bottom-up VA is largely involuntary and influenced by the intrinsic features of visual stimuli, collectively referred to as their saliency. Saliency describes the property of an

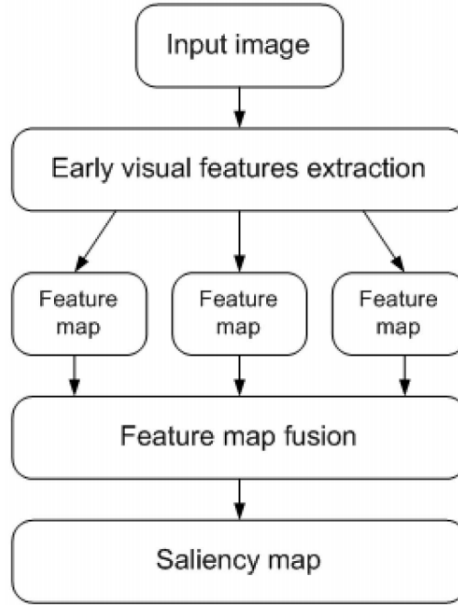


Figure 5.1: Saliency framework proposed by Koch and Ullman (1987) consisting of parallel extraction of low-level features across multiple channels and their fusion into a final saliency map.

object ‘standing out’ from its neighbours. Saliency and bottom-up VA are rooted in evolutionary theory and are linked to the facilitation of survival, by deploying VA to the most relevant information in a scene and responding to sudden threats, such as a predator (Borji, Sihite and Itti, 2013). This suggests that saliency is computed rapidly across the entire visual field (25-50 ms per item), prior to deployment of volitional attention, which often requires additional eye movements (Itti and Koch, 2001). The bottom-up saliency framework, originally formalised by Koch and Ullman (1987), poses saliency as a weighted combination of different feature maps derived from low-level features of the stimulus by the HVS (see Fig. 5.1). This approach to feature extraction and aggregation is strongly rooted in Marr’s theoretical computational vision models, and also underpins contemporary computer vision techniques, including convolutional neural networks, which rely on hierarchical feature extraction and fusion for many common tasks. In the context of compositing, a similar framework can be used for prediction of visual realism, whereby various inconsistencies in an input image are extracted and weighted into a fused feature map indicating local realism estimates. The challenge in both the VA and realism scenario is the specification of features to extract.

In contrast, top-down VA is driven by the visual task at hand and consequently volitional deployment of attention. As this process is controlled by higher cognitive areas, it requires additional effort, compared to bottom-up VA (Itti and Koch, 2001). Top-down VA is employed in tasks such as *visual search*, where the goal is to find objects with some pre-specified properties. This allows for the HVS to attend to objects relevant to the task and disregard any irrelevant information. Top-down VA has a modulating effect on bottom-up

attention, for example by affecting the relative weightings of features contributing to the computed saliency map, based on the demands of a given task (Treue and Trujillo, 1999). Verbal comments of observers from the experiment presented in Chapter 4 suggest that top-down attention plays a key role during the assessment of visual realism. Specifically, observers reported comparing the appearance of the target object to objects with similar appearance or illumination. If observers perform some inference based on relevant scene content, the selection of relevant targets would require deployment of volitional attention.

Both top-down and bottom-up VA mechanisms tend to operate in parallel in everyday scenarios, whereby attention is both modulated by the intrinsic properties of objects, as well as top-down properties such as memory, task and context.

The Feature Integration Theory & Visual Search

Treisman and Gelade (1980), in her seminal work on focused visual attention, proposed the *feature integration theory*, which aimed to reconcile the existence of the *visual pop-out effect* (Wang, Cavanagh and Green, 1994), which implied the existence of pre-attentive feature extraction and grouping, with the deployment of focused volitional attention. Treisman’s feature integration theory posits that the HVS extracts multiple separable, low-level features globally from a visual stimulus, before deploying focal attention to bind these features together into a coherent percept. Based on this, she highlights two distinct visual search mechanisms: feature search and conjunctive search. Feature search is the pre-attentive feature extraction stage, responsible for the pop-out effect. It can be performed rapidly, but only in scenarios where the target can be defined by a single feature. In other scenarios, overt VA is deployed in a serial manner, fusing and evaluating combinations of multiple features, before locating and identifying the target.

Accordingly, the authors showed that individual visual features can be detected and identified in a parallel, pre-attentive fashion, without requirement for their explicit localisation. However, conjunctions of features rely on application of serial focal attention and cannot be identified without prior localisation. Analysis of real world scenes, in context of a visual realism analysis task, relies on the process of visual search. For example, in order to reliably rate the realism of a scene, observers must perform visual search to find evidence of manipulation or editing. In some scenarios, particularly ones where the distortion is really obvious, this may rely on feature search and the pop-out effect. However, for more realistic composites (or more subtle distortions), conjunction search may be required to identify and localise the ‘odd’ object.

Local and Global VA

VA can also be discussed in the context of scale of visual information attended. There is evidence that VA is deployed at two distinct scales: *global* and *local* (Liechty, Pieters and Wedel, 2003). Global VA acts as an initial guiding process, selecting components of the visual field for local VA, which in turn extracts detailed information from these relevant

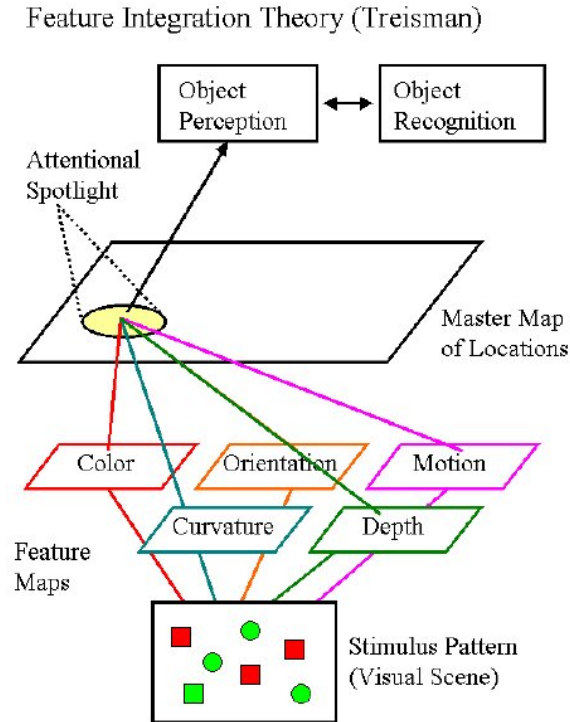


Figure 5.2: Overview of Treisman’s Feature Integration Theory. Image courtesy of Treisman and Gelade (1980)

regions. These two distinct mechanisms of VA - orienting and resolving - drive the process of information selection and processing (LaBerge, 1998). Visual information extracted by global VA processes is not only utilised for the purpose of orienting, but has been shown to play a significant role in scene recognition (Biederman, Mezzanotte and Rabinowitz, 1982). Computational models of this process have been proposed by Oliva (2005). Local VA is associated with top-down processes, which have been shown to significantly affect where attention is deployed. For example, Ninassi et al. (2006) show that observers utilise different visual strategies for assessing different types of image distortions, as well as when performing different visual tasks. In the context of compositing and visual realism, it is possible that significant distortions may trigger bottom-up, or global mechanisms, attracting overt attention to the distorted region for further analysis. On the other hand, minor distortions may require deployment of local, or top-down processes in order to extract further task-relevant information. This transition between covert and overt attention as a function of distortion severity can be motivated by much of the literature discussed in Chapter 2, where some physical or semantic inconsistencies may go unnoticed, while others are readily detected.

5.2.2 Modelling VA

Top-down VA is commonly modelled based on statistical analysis of empirical human gaze data, commonly recorded using specialist hardware (Yarbus, 1967). Relevant literature

on eye tracking is discussed in Section 5.2.4. In contrast, bottom-up VA is often rooted in computational models of the HVS, or its particular mechanisms, such as the parallel feature extraction and fusion framework proposed by Koch and Ullman (1987). Many approaches to the task of computational VA modelling have been suggested, many focusing on bottom-up attention and saliency, due to their relative simplicity and suitability for computational modelling, based on the feature integration theory of attention (Treisman and Gelade, 1980). The goal of such models is to accurately predict highly-probable locations of human fixations in an image, thus pointing out attention-grabbing regions or objects in an image, and representing this information as a 2-D probability map. This makes the concept of attention or saliency maps attractive in the computational modelling and image processing domains, as the saliency map can be directly used to spatially weight the underlying image features.

Saliency maps can be generated adopting either a local approach, using centre-surround differences computed in local regions, or a global one, utilising features of the entire image. While being biologically-plausible, both approaches come with their distinct advantages and associated problems. For example, local methods tend to overestimate the saliency of object edges and high-frequency image content, while missing larger salient regions, while global methods tend to work well for larger objects, but have trouble dealing with highly-textured regions. Itti, Koch and Niebur (1998) in their seminal work proposed a biologically-inspired bottom-up saliency-based model of VA. This model, given an input image, generates a corresponding saliency map through successive stages of colour, intensity and orientation feature extraction, centre-surround differencing and linear integration of these features at multiple scales. This model served as a foundation for multiple further developments, such as the graph-based saliency model of Harel, Koch and Perona (2007) or the conditional random field-based approach of feature combination put forward by Liu et al. (2011).

More recently, with the resurgence of deep learning, deep convolutional neural networks (DCNNs) have been adopted to model saliency, achieving state-of-the-art results, by performing end-to-end feature learning, extraction, integration and saliency prediction (Borji, 2019). Despite this, human performance has not yet been matched. Examples of this are plentiful, including Wang et al. (2015), Wang and Shen (2017), Zhao et al. (2015), Pan et al. (2016), Pan et al. (2017) to name a few. The vast majority of deep learning-based saliency prediction models learn the function mapping from image to saliency map directly from empirical data. Differences between them are mostly rooted in training data, model architecture and hyperparameters, with the general goal of approximating the function performed by the observer HVS unchanged.

In addition to modelling saliency for static displays, such as images, temporal saliency models have also been proposed, aiming to incorporate aspects such as object movement into saliency predictions. For example, Marat et al. (2009) create both a static and temporal saliency map from video input and combine both into a spatio-temporal saliency

map. Bak et al. (2017) build on this approach by using deep convolutional neural networks to both extract and combine the spatial and temporal saliency maps. While temporal saliency models highlight an interesting future research direction (e.g. modelling the entire realism assessment process), existing models would be challenging to adapt to static displays, without disregarding the temporal aspect.

5.2.3 Applications of VA Models

VA modelling plays a crucial role in perceptual models, allowing for a relevance-based ranking of visual stimuli and a perceptual weighting of image content to be computed. This is particularly important in tasks where human performance needs to be matched closely, such as image quality assessment or visual realism prediction. This is further reinforced by the fact that allocation of VA has been shown to affect perceptual thresholds (Orquin and Loose, 2013). Accordingly, there is evidence that distortions or disparities in salient regions are more likely to contribute to a lower subjective quality score than those in non-salient regions (Engelke et al., 2010). However, this effect is highly dependent on context and task. For example, Ninassi et al. (2007) evaluated the impact of attention-based spatial pooling functions on image quality scores and found that results were not conclusive, suggesting nonlinear relationships between attention and distortion magnitude. This evidence suggests that the interplay between attention, subjective rankings and distortion type is highly variable as a function of image content, distortion and task. This has significant implications for modelling of visual realism, which in turn depends on multiple image features, as well as viewing contexts (e.g. free viewing or task-based viewing). In order to accurately model this process in the context of composite realism, it is important to verify whether the visual strategy deployed by observers changes as a function of distortion type and whether prior knowledge about the distortion location can change subjective realism scores. In the study presented in Chapter 4 the effects of VA were controlled for by provision of explicit prior knowledge of the identity of the composited object to observers. In practical scenarios, this information is scarcely available, accordingly the detection of composite artefacts may be affected by deployment of VA. Any computational model of VA, particularly ones used in perceptually-relevant applications, should be relevant to the task at hand. This is both due to the difficulty of VA modelling, as well as the impact of a task on VA allocation in humans. No prior studies have investigated visual attention in the context of image composite realism. Accordingly, an empirical study of overt VA could shed light on deployment of visual strategies and illustrate the impact of factors such as transformation type, or prior object knowledge on resulting visual realism ratings.

5.2.4 Gaze Tracking for VA Modelling

Tracking eye movements offers a reliable proxy for measuring overt VA. Human eye movements can provide objective information complimentary to conventional subjective ratings, such as questionnaires or rating scales (Elhelw et al., 2008). As eye movements are

paramount to acquisition of visual information while performing cognitive tasks, studying how they are deployed can reveal visual strategy and features relied upon during the completion of a task (Duchowski, 2002).

Eye movements are commonly analysed in terms of two canonical classes of events: *fixations* and *saccades*. Fixations are short pauses over informative regions of stimuli and are interspersed by saccades - rapid movements from one region of the stimulus to another. In practice, eye tracking is commonly performed through measurement of eye movements relative to some external stimulus, commonly a display or real world scene. This is accomplished using head-mounted or desktop eye trackers, which capture images of the observer's eyes and recover the position and orientation of the pupils. This information is then projected into the coordinate system of the stimulus. Thus, the instantaneous gaze position can be recorded many times per second and spatially related to the stimulus. The raw sampled gaze positions are then classified into fixations and saccades. Multiple approaches to this process exist and rely on exploiting different features of the gaze data. For example, velocity-based algorithms rely on the fact that fixation points have low velocities compared to saccades. Dispersion-based algorithms evaluate the spread of gaze data under the assumption that fixation data points tend to be spatially clustered together, while saccades are spatially spread out. Salvucci and Goldberg (2000) provide a taxonomy and overview of existing fixation detection methods, while Komogortsev et al. (2010) provide a set of performance metrics for evaluation of different fixation detection algorithms. Non-parametric approaches have also been proposed, removing the need for manual parameter tuning (König and Buffalo, 2014).

Examples of the use of gaze data in the assessment of visual strategy and attention exist in both free-viewing conditions (Yarbus, 1967), as well as specific tasks such as reading (Reichle et al., 1998), visual search (Rajashekar, Cormack and Bovik, 2004), objective image quality metrics (Ninassi et al., 2007), decision-making (Orquin and Loose, 2013), scene perception (Henderson and Hollingworth, 1999) and subjective quality evaluation (Venkatesh and Sen-ching, 2010). However, there are few cases of gaze data used in subjective evaluation of visual realism. Zangemeister, Sherman and Stark (1995) used eye tracking to analyse visual strategies when viewing abstract and realistic art. Ninassi et al. (2006) used objective eye metrics to study the impact of task on VA in subjective image quality assessment through comparing empirical fixation maps. Elhelw et al. (2008), studied the impact of different image features on the perceived realism of real and synthetic bronchoscopy images. Finally, Vu et al. Vu, Larson and Chandler (2008) assessed the impact of common global image distortions, such as blurring, noise, packet loss and JPEG compression artefacts on fixation patterns.

5.2.5 Visual Tasks & Gaze Metrics

Commonly, fixations and saccades are spatially and/or temporally aggregated to reveal the distribution of overt attention. Additionally, properties of individual fixations or saccades

can be aggregated and compared. The specific metrics used for analysis of gaze data differ depending on the experimental design. A wide range of fixation metrics have been used for a number of different experimental scenarios and visual tasks (Rayner, 1998; Gegenfurtner, Lehtinen and Säljö, 2011; Bylinskii et al., 2015; Sharafi et al., 2015).

Fixation Count / Ratio

Fixation count refers to the number of fixations in a given region of the stimulus, commonly calculated within a fixed time frame. Fixation counts are commonly related to the informativeness of a particular region in an image and positively correlate with visual effort (Rayner, 1998). Fixation counts can also be normalised and represented as a ratio of all fixations in order to compare relative informativeness of different regions in the stimulus.

Fixation Duration

Fixation duration describes the time needed to analyse a given stimulus. Longer fixation durations correlate with more visual effort required for a given region of the stimulus (Rayner, 1998). This is commonly related to the difficulty of a given visual task. For example, Levy-Schoen (2017) found that mean fixation durations were shorter for silent reading, compared to reading aloud. This is also true for visual search tasks. Redundancy or predictability (Nattkemper and Prinz, 1984), as well as the overall complexity of a visual search task, were shown to influence fixation duration Moffitt (1980).

Fixations on Target

In tasks which rely on a *target* region in the stimulus, such as visual search, early fixations should be interpreted differently from later ones. This is due to the fact that early fixations tend to be driven by bottom-up VA mechanisms, as opposed to later fixations, which result from task-related top-down attention processes (Jacob and Karn, 2003). In addition to this, the duration of early fixations on target objects has been shown to be a positive indicator of their task-related information content (Henderson and Hollingworth, 1998).

Area of Interest

Grouping fixations by semantic regions they fall on (e.g. objects in an image) allows for area of interest analysis. Gaze maps can also be generated and compared by aggregating fixations over time and representing them as 2-D distributions. These can be compared across scenes or observers using common distribution similarity metrics in order to identify significant differences between areas of interest.

Observer Metrics

When performing gaze analysis across groups, measures of intra- and inter-group consistency allow for distinctions between bottom-up and top-down processes.

Marius’t Hart et al. (2009) suggests that any bottom-up model of attention must ensure high inter-observer consistency in order to make saliency predictions. Thus, observations of high consistency between observers may be an indicator of the precedence of bottom up processes over top-down ones. This is particularly important when taking the magnitude of distortions into consideration. Large magnitude distortions in images are likely to attract attention more consistently, compared to subtle ones. Similarly, top-down factors, such as prior knowledge or instructions, can also increase this measure through constraining the options presented to observers.

Intra-observer measures allow for such evaluation between individual stimuli presented to a single observer. This allows for the assessment of variations in stimuli on observer gaze statistics.

5.2.6 Application of Gaze Metrics

While a considerable proportion of gaze tracking research has focused on analysis of reading, efforts to understand eye movements in visual search and scene understanding (Rayner, 1998), both of which are closely related to the task of image composite realism assessment.

Scene Analysis

Prior research (Rayner, 1998) suggests that during scene analysis, observers abstract the gist of the scene within the first few fixations. Following this, Yarbus (1967) points out that observers continue to extract relevant information from the scene through deployment of additional fixations. Objects that are informative, important or appear out-of-place in a scene are often fixated for longer (Antes, 1974; Friedman, 1979; Loftus and Mackworth, 1978), suggesting that severe distortions in images, or image composites, may attract observer attention. Furthermore, scene context can sometimes have an immediate effect on object processing. For example, Boyce and Pollatsek (1992) found that background information extracted from the first and second fixations aids object identification. This indicates that immediate scene context, as well as the contrast between the appearance of the object and its local neighbourhood, should be considered in visual search and discrimination tasks.

Visual Search

Similar to the scenario of scene analysis, in visual search, the properties of the task and stimulus tend to affect the specific pattern of VA deployment. For complex visual search, such as the search for complex features in a natural image, Zelinsky et al. (1997) showed that observers have a central bias, deploying initial fixations to the centre of the display, before recursively deploying fixations to smaller groups of objects in the display.

5.2.7 Summary

VA is a fundamental mechanism of human visual perception, particularly when visual stimuli represent natural, real-world scenes. Aside from semantics and scene content, VA is heavily modulated by task, where bottom-up processes co occur with top-down, task-dependent and volitional shifts of attention. Prior research has shown that the eye tracking paradigm, along with a set of gaze metrics, can be an accurate tool for measurement of VA allocation in a range of scenarios, including ones closely related to assessment of visual realism in image composites, such as visual search and scene analysis. While many general models of VA exist in the literature, none of them have been designed with image composite assessment and associated peculiarities in mind, making them challenging to apply without extensive evaluation and modification. Furthermore, Chapter 4 illustrated that top-down processes play a clear part in this task, suggesting that saliency models may not explain observer behaviour sufficiently in this context. In order to understand the potential impact of VA of the task of composite image assessment, as well as the generalised psychometric models of visual realism proposed in the previous chapter, a subjective gaze tracking study is proposed.

5.3 Methodology

5.3.1 Overview

The study presented here evaluates how *local transformation type* and *prior object knowledge* affect observer gaze metrics and subjective realism ratings. This is accomplished through statistical comparisons of fixation data collected from 4 groups of 15 observers performing a visual realism rating task using stimuli described in Chapter 4. Under this paradigm, significant differences between group behaviour can be highlighted and through extraction and comparison of attention maps, areas of interest can be identified and compared across groups, providing insight about the visual information most relevant to the task.

In Chapter 4, three different transformation types were evaluated: exposure, contrast and CCT shifts. Of these, contrast transformations were found most challenging to reliably detect by observers. In order to reliably compare between distributions of fixations for multiple conditions, contrast transformations are excluded in this study. Instead, both exposure and CCT transformations are considered, particularly due to the similarity of their lapse rate estimates, indicating comparably low average observer error rates, compared to contrast. In addition, the impact of prior object knowledge, and, by proxy, task-relevant instructions likely to guide VA is also assessed. Accordingly, a between-groups 2×2 factorial design is adopted in order to measure the impact of two independent categorical variables:

- **local transformation type:** exposure (**E**) and correlated colour temperature (**C**)

- **prior object knowledge:** unknown (**U**) and known (**K**)

The dependent variables are represented by various fixation metrics, derived from the collected gaze data. They are:

- fixation count
- fixation duration
- time to first fixation on target object
- area of interest similarity
- inter-observer consistency
- response time
- realism rating

and are introduced in Section 5.2.5 and discussed in the context of the experiment in Section 5.3.6.

Using the 2×2 factors described above, 4 experimental conditions are defined (see Table 5.1): **EU** (exposure, unknown), **EK** (exposure, known), **CU** (CCT, unknown), **CK** (CCT, known). To ensure that observers view images in a natural context, the 2AFC procedure adopted in Chapter 4 cannot be used, since observers are able to see the processed and unprocessed image at the same time. Instead, the experimental procedure selected is an adaptation of the *double-stimulus impairment scale* (DSIS) method (ITU, 2002), where, in each trial, observers are requested to indicate if an object-scene combination affected by a particular local transformation appears realistic or unrealistic, given a reference image displayed prior to the test image. In this scenario, distribution of VA for the test image is task-specific and not corrupted by fixations intended to compare between multiple images. If the transformation type or prior object knowledge significantly affect perceived realism or assessment strategy, this should be reflected in group realism ratings and overt attention distributions, respectively.

In the wider context of this thesis, the main application of this study is in assessing whether deployment of visual attention can significantly change resulting subjective realism judgments. A positive answer to this question would suggest that VA modelling be explicitly incorporated in any systems attempting to predict human realism responses. Moreover, in this scenario, realism responses would likely depend on the content presented to the viewer immediately prior to the realism judgment. Conversely, a negative answer to this question allows any downstream modelling of human realism perception to safely forgo explicit modelling of human visual attention.

		Prior Object Knowledge	
		Unknown	Known
Transformation Type	Exposure	Group A (EU)	Group B (EK)
	CCT	Group C (CU)	Group D (CK)

Table 5.1: Experiment design, showing 2×2 factorial design and assigned observer groups.

5.3.2 Stimuli

To generate the composite images serving as the stimuli for the experiment, the synthetic composite approach used in Chapter 4 is adopted. Specifically, 33 images with segmented objects are selected from a subset of the SUN Dataset used in Chapter 4. The images are selected to cover a range of object types and luminance values, according to the mean luminance of the segmented objects, calculated in the CIE $L^*a^*b^*$ colour space. The selection is made such that in each image, the segmented objects occupy no more than $1/3$ of the total image area. The horizontal resolution of the images is normalised to 600 pixels (px), preserving the aspect ratio. At a 65 cm viewing distance, 37 px on the screen correspond to 1° visual angle (VAn). Examples of these images can be seen in Chapter 4, e.g. Figures 4.15 through 4.20.

For each base image, two synthetic composites are generated by applying a local exposure and CCT transformation to the segmented object. Based on the lower performance on contrast shifts observed in Chapter 4, contrast transformations are omitted in this study. The image-magnitude combinations are fixed across all conditions, meaning that for each base image, the magnitude of the exposure or CCT transformation is kept the same across both object knowledge conditions. As in Section 4.3.4 exposure transformations are implemented using a scaling of the V channel of HSV colour space, whereas CCT transformations are implemented using an additive offset in the perceptually uniform *mired* space. The transformations applied to the objects are varied in terms of magnitude: exposure is scaled in 11 logarithmically spaced steps between .3162 and 3.162, corresponding to a range of -1.661 to 1.661 in \log_2 domain. CCT is offset in 11 increments of 40 mired, between -200 and 200 mired. The order of relative offset intensities is kept the same for exposure and CCT. Examples of offsets applied to an image can be seen in Figure 5.3.

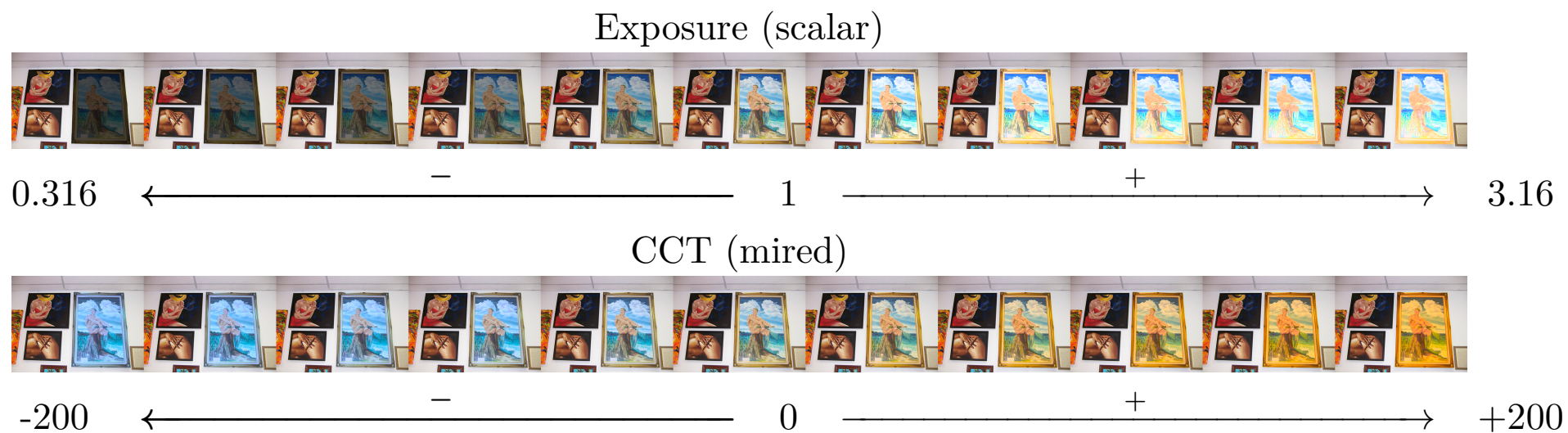


Figure 5.3: Offsets applied to segmented objects in test images. Top row: exposure offsets (scalar multiplication); Bottom row: CCT offsets in mired (subtraction / addition)

5.3.3 Observers

Sixty (60) observers, recruited from a population of university staff and students, are randomly assigned into 4 groups of 15 observers (see Table 5.1). All observers are volunteers and are not rewarded for their participation in the study. The following groups are compiled: *Group A* (condition EU) with a mean age of 26.00 ($SD = 5.76$) 7 females, *Group B* (condition EK) with a mean age of 25.93 ($SD = 4.32$), 7 females, *Group C* (condition CU) with a mean age of 28.47 ($SD = 4.81$) 6 females and *Group D* (condition CK) with a mean age of 32.27 ($SD = 8.15$) 7 females. All observers are checked for normal or corrected-to-normal vision and normal colour vision, using a SNELLEN chart and Ishihara test. Each observer is requested to provide consent to take part in the experiment and is naïve to its purpose.

5.3.4 Apparatus

Display

Images are displayed on a 22" 60 Hz Iiyama ProLite B2280HS LED monitor, calibrated to sRGB colour space using an X-Rite i1 Display Pro calibrator. The monitor is placed in an evenly illuminated room, and the calibration is corrected for both the chromaticity and intensity of the ambient illumination. The maximum measured luminance level of the display is 214 cd/m^2 , while the black luminance is $.375 \text{ cd/m}^2$. When displayed, the images occupy $11.8^\circ \times 7.9^\circ$ VAn.

Eye tracker

A Tobii X1 Light eye tracker is fixed below the display at a distance of 65 cm from the observers' head (see Fig. 5.4), as recommended by the manufacturer. Average binocular accuracy, as reported by the manufacturer, is $.4^\circ$ VAn and an average precision is $.2^\circ$ VAn at the selected viewing distance. Its typical sampling rates fall between 28-32 Hz. The eye tracker compensates for head movements of up to 44 cm horizontally and 32 cm vertically, removing the need for a chin rest. The device is recalibrated for each observer, following the manufacturer's recommendations (Tobii, 2012).

5.3.5 Procedure

Preparation

Observers are first asked to familiarise themselves with the instructions and shown examples of the reference (original unmodified - see Fig. 5.5a) and test images (processed object - see Fig. 5.5b). Observers are then given an opportunity to ask questions, before commencing the experiment.

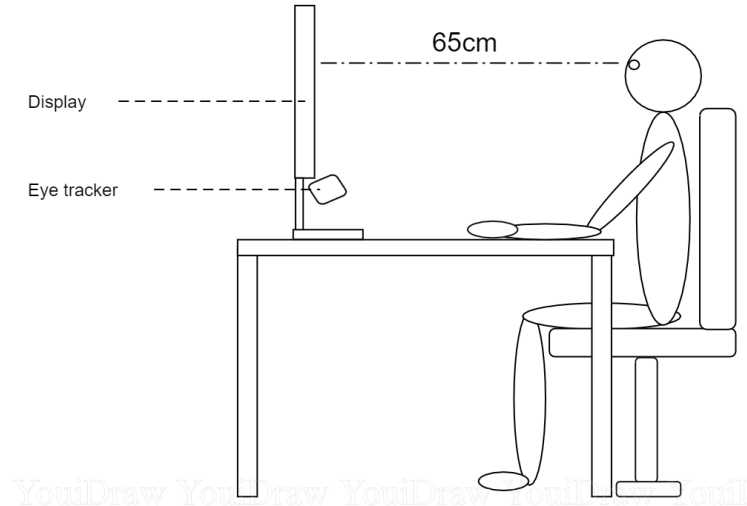


Figure 5.4: Experimental setup: Observers are positioned at a distance of 65cm from the display. The eye tracker is mounted at the bottom of the display and re-calibrated for each new observer. During the experiment, observers use a mouse to provide their responses.

Trials

The experiment consists of 33 trials, broken into 3 sessions with short breaks in between. During each trial, observers first see a reference image, which they are instructed to analyse. In the case of conditions *EK* (*Group B*) and *CK* (*Group D*), a binary mask is also displayed next to the reference image, revealing the identity of the foreground object of interest (see Fig. 5.5a). This stimulus is displayed for 10 seconds, followed by a 3-second middle grey screen to ensure change blindness (Wolfe, 2000). Next, the test image is displayed (see Fig. 5.5b), which contains the effects of a local exposure or CCT transformation applied to the segmented object. The observers' task is to provide a binary realism rating of the test image with respect to the reference image. Observers have 10 seconds to analyse the test image and click the button corresponding to their chosen answer. They are also requested to respond as quickly and accurately as possible. This procedure is repeated for each observer. Listing 5.1 shows the verbatim instructions provided to all observers.



(a) Reference screen (binary mask only shown to observers from groups B and D)



(b) Test screen

Figure 5.5: Illustration of reference and test stimuli and associated interface presented to observers during the experiment. a) Reference screen, containing an unmodified version of an image and, in the cases of groups B and D, a binary mask indicating the target object. b) Test screen, containing a transformed version of the image from the reference screen. The transformation is only applied in the target area indicated by the binary mask. The binary mask is only displayed to groups B and D, while groups A and C only see the original and modified image. Observers are asked to provide a binary rating of the realism of the test image in comparison to the reference by clicking the corresponding button.

```

Hello and thank you for taking part in this experiment.

The task:

1. In this experiment you will be shown 3 series of 11 images.

2. During each trial you will see two instances of each image.

3. FIRST, you will see a reference image - make sure you familiarise yourself
with this image. You will have 10 seconds to do this.

--
Groups B and D only:
(The location of the target object will be indicated by a black & white image.)
--

4. After this you will see a grey screen for 3 seconds.

5. Then, the SECOND version of the same image will be displayed.
One of the objects in this image will be modified/adjusted in some way.

--
Groups B and D only:
(This is the object indicated by the black & white image in the previous screen.)
--

6. Your task is to decide whether the second image appears REALISTIC or UNREALISTIC.
You can do this by pressing a one of the two buttons using the mouse.
You will have 10 seconds to do this.

7. IMPORTANT:
The speed and accuracy of your response for the SECOND image will be measured.
Please respond as quickly and accurately as you can.

During the experiment, the position of your gaze will be tracked.
Please stay as static as possible and do not move between trials.

```

Listing 5.1: The instructions presented to each observer in the experiment.

5.3.6 Analysis

Fixation Extraction and Fixation Maps

Fixations and saccades are extracted from the raw eye position data using the ClusterFix package for MATLAB (König and Buffalo, 2014). Resulting fixation locations are then aggregated into fixation maps (sparse 2D histograms) \mathbf{F}_o , for each image-observer combination. Only fixation data from the image region is used in this process. Fixations falling outside the reference and test image regions (such as the user interface) are rejected. Joint fixation maps \mathbf{F}_{joint} are then generated by normalising and averaging fixations for

each image across all observers from a single condition:

$$\mathbf{F}_{joint} = \frac{1}{N} \sum_{o=1}^N \frac{\mathbf{F}_o}{\sum_{i=1}^H \sum_{j=1}^W F_{ij}^o} \quad (5.1)$$

Here, F_{ij}^o denotes element at row i and column j of fixation map \mathbf{F}_o , H and W denote the height and width of the image, and Each joint histogram is normalised by the sum of its elements N , in order to avoid biasing the joint fixation maps towards observers with higher fixation counts. As such, each bin of the fixation map represents a proportion of task time a location was fixated by an observer.

Eye Movement Metrics

The following fixation metrics are used:

- fixation count (Fc)
- fixation duration (Fd)
- time to first fixation on object ($TFFO$)
- duration of first fixation on object ($DFFO$)

As discussed in Section 5.2.5 Fc correlates positively with the amount of information to be attended and task difficulty and Fd relates to the usefulness of particular regions to task completion and overall difficulty of information extraction. A distinction must be made between early fixations, driven by bottom-up VA mechanisms, and later fixations, driven by top-down mechanisms. Accordingly, shorter $TFFO$ can be an indicator of an object attracting bottom up VA (Jacob and Karn, 2003) and longer $DFFO$ can point to an object’s task-relevant semantic informativeness (Henderson and Hollingworth, 1998). Furthermore, Fd can be affected by scene context, e.g. objects that do not belong in the scene tend to attract longer fixations than objects that do (Rayner, 1998). These properties allow for reasonable interpretation of significant differences between groups in terms of visual effort and task-relative usefulness of image features.

Fixation Map Metrics

To assess similarity between observers’ fixation distributions within one condition, the inter-observer consistency (IOC) measure is used (Le Meur and Baccino, 2013), specifically the “one against all” approach. This compares the fixation map of each observer against a joint fixation map of all other observers using a similarity metric. The area of interest similarity ($AOIS$) measure is also adopted, encoding similarity between joint fixation maps across experimental conditions.

Many standard methods for calculating the similarity between joint fixation maps exist. Methods such as Area Under the Receiver Operator Characteristic curve, Kullback-Leibler



Figure 5.6: Visualisation of a heatmap generated from the distribution of individual fixations, marked with \times .

divergence, or Earth Mover’s Distance (see Riche et al. (2013) for a review). Many of these methods present the problem of comparing predicted saliency maps against ground-truth fixation maps as a classification evaluation problem. In the scenario presented here, there is no objective ground truth, since both fixation maps being compared are empirical, and the goal of comparing them is to obtain interpretable measures of similarity. Accordingly, for both *IOC* and *AOIS*, the similarity score (*SS*) is used, as recommended by Riche et al. (2013). A key benefit of this metrics is that it is bound between 0 and 1, allowing for straightforward interpretation of the results, as opposed to unbounded and asymmetrical methods such as KL-divergence. The similarity score approach computes the sum of the minima between each point of two probability distributions:

$$SS(P, R) = \sum_{i,j} \min(P_{i,j}, R_{i,j}) \quad \text{where} \quad \sum_{i,j} P_{i,j} = \sum_{i,j} R_{i,j} = 1.0 \quad (5.2)$$

P and R represent discrete 2D probability distributions (PDs). We convert discrete fixation maps to PDs by placing a Gaussian with $\sigma = 1^\circ VAn$ at the location of each fixation in order to model uncertainty in viewing location caused by the accuracy and precision of the eye tracker, as in Le Meur and Baccino (2013). See Figure 5.6 for an example of a heat map visualisation of fixation distributions, along with the original fixations, marked as red ‘x’ symbols.

Statistical Measures

Non-parametric Monte-Carlo-based methods, proposed by Efron and Tibshirani (1994), are used to analyse gaze metric distributions. Specifically, bootstrapping is used both to estimate the means/medians of the eye movement metrics from the empirical samples, as well as calculate their 95% confidence intervals (CIs), standard errors and bias. The bias-corrected and accelerated (BCa) method is used to calculate CIs (Owen, 1988). Furthermore, bootstrapping techniques are also used to compare group differences and effect sizes. Fisher’s permutation test Odén and Wedel (1975) is adopted as a means for testing the statistical significance of differences between groups. The chosen statistic for this procedure is the difference of means ($\bar{x}_1 - \bar{x}_2$), unless the empirical data distribution is heavily skewed - in such cases the difference of medians is used, as this is less sensitive to high variance. The number of simulated samples for the bootstrap procedures and permutation tests is 5000. When computing correlations across realism responses, Pearson correlation is used.

5.3.7 Qualitative Analysis of Fixation Maps

In addition to the statistical evaluation presented in preceding sections, fixation maps are evaluated visually, both for purposes of quality assurance, but also in order to perform qualitative assessment of attention patterns. A simple methodology is developed for this task: each heatmap is assessed to find the most attended objects, and to evaluate how attention is distributed throughout the semantic content of the scene. This process provides an additional means of hypothesis verification, as well as providing a source of information about potential influences of scene semantics on particular gaze statistics.

5.3.8 Hypotheses

To evaluate whether the factors under test affect deployment of observer attention and subsequent realism responses, a selection of relevant gaze metrics are compared between conditions. Based on the results from Chapter 4 and a review of visual attention literature, the following hypotheses are evaluated:

Eye Movement Metrics:

H1: Fixation counts (F_c) for groups with prior object knowledge will be lower.

H2: Fixation durations (F_d) for groups with prior object knowledge will be higher.

H3: Time to first fixation on target object ($TFFO$) will be shorter for groups with prior object knowledge.

H4: Duration of first fixation on object ($DFFO$) will be longer for groups with prior object knowledge

Fixation Map Metrics:

H5: Inter-Observer Consistency (*IOC*) will not be affected by either of the factors

H6: Area of Interest Similarity (*AOIS*) will be highest for groups sharing object knowledge priors

Task Performance:

H7: Realism ratings will be lower for groups with prior object knowledge

H8: Response times will be shorter for groups with prior object knowledge

Justification

Overall, it is expected that prior object knowledge is likely to make the overall task easier, by reducing the need for observers to perform visual search, providing them with task-relevant information and reducing uncertainty. Consequently, groups where prior object knowledge is provided may be able to notice more visual inconsistencies, compared to groups without this information, resulting in lower average realism scores (*H7*). Hypothesis *H1* indicates that this reduction in uncertainty and visual search will in turn reduce the number of fixations required for observers to make a decision, also shortening overall response times, as indicated by *H8*. As fixations of observers in these groups are likely to land on task-relevant regions, *H2* accordingly predicts that they are likely to be on average longer. Accordingly, *H3* and *H4* suggest that observers will fixate on the target object sooner and for a longer duration when they are aware of its identity. No significant differences in inter-observer consistency are expected between groups, as indicated by *H5*. However, the locations attended by observers may be more similar for groups where the target object is made explicit, as indicated by *H6*, since observers' attention is guided there by the binary mask in the reference screen. Accordingly, groups with prior object knowledge may contribute to higher measures of *IOC*.

As no prior studies have assessed perceptual differences between transformation features in this context, all above hypotheses are null with respect to the second factor - transformation type. This is with exception of *H7*, as following the results in Chapter 4, the change in transformation type is expected to impact realism judgments.

5.4 Results

This section presents the results of the experiment. The following subsections detail the results for each one of the dependant variables in the presented study. As an overview, Figure 5.8 illustrates the distributions of gaze metrics for each of the conditions studied. Specifically, bootstrapped means along with their 95% confidence intervals are shown for each combination of gaze metric and experimental group. Additionally, Figure 5.9 shows

the means and 95% confidence intervals of the difference statistics used when performing group comparisons, described in Section 5.3.6.

5.4.1 Realism Ratings

Contrary to *H7*, the factor of prior object knowledge was not found to have a significant impact on resulting realism ratings. Figure 5.7 shows plots of mean realism responses per transformation magnitude value, where each mean realism response value is an average of binomial responses across all images with the same transformation magnitude. Error bars plot 95% binomial CIs for each realism rating in the experimental group. It can be seen that the confidence intervals for each measured transformation magnitude overlap significantly in the case of both transformation features.

Fisher’s exact test indicated that, in the case of each transformation type, prior object knowledge had no significant impact on group realism ratings ($p > 0.05$). While such a direct comparison of realism ratings across transformation type is not possible, due to the exposure and CCT transformation magnitude scales not being perceptually aligned, a moderate correlation was observed between conditions EU and CU ($r = 0.57, p < 0.05$), as well as EK and CK ($r = 0.63, p < 0.05$). Stronger correlations were observed between condition pairs EU and EK ($r = 0.91, p < 0.05$) and CU and CK ($r = 0.81, p < 0.05$).

5.4.2 Fixation Counts

Overall, conditions where object identity was known a priori received significantly lower mean F_c , as indicated by Fisher’s permutation test ($p < 0.05$). No significant F_c differences were found between groups affected by exposure and CCT transformations, when the object was known. When prior object knowledge was not provided, however, significantly higher mean F_c values were recorded for CCT, compared to exposure transformations.

This suggests that overall, CCT transformations in the range evaluated may have been more challenging to detect, compared to exposure transformations, necessitating extended visual search. See Figures 5.8a and 5.9a for an illustration of these statistics and their respective effect sizes. These results indicate that visual search plays a significant role in

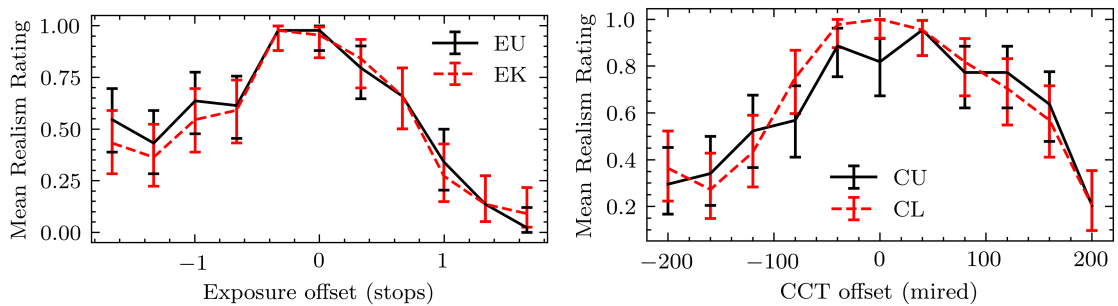


Figure 5.7: Realism responses averaged for feature offset values across image sets for exposure (left) and CCT (right). Line styles indicate object location conditions. Error bars indicate 95% binomial confidence intervals.

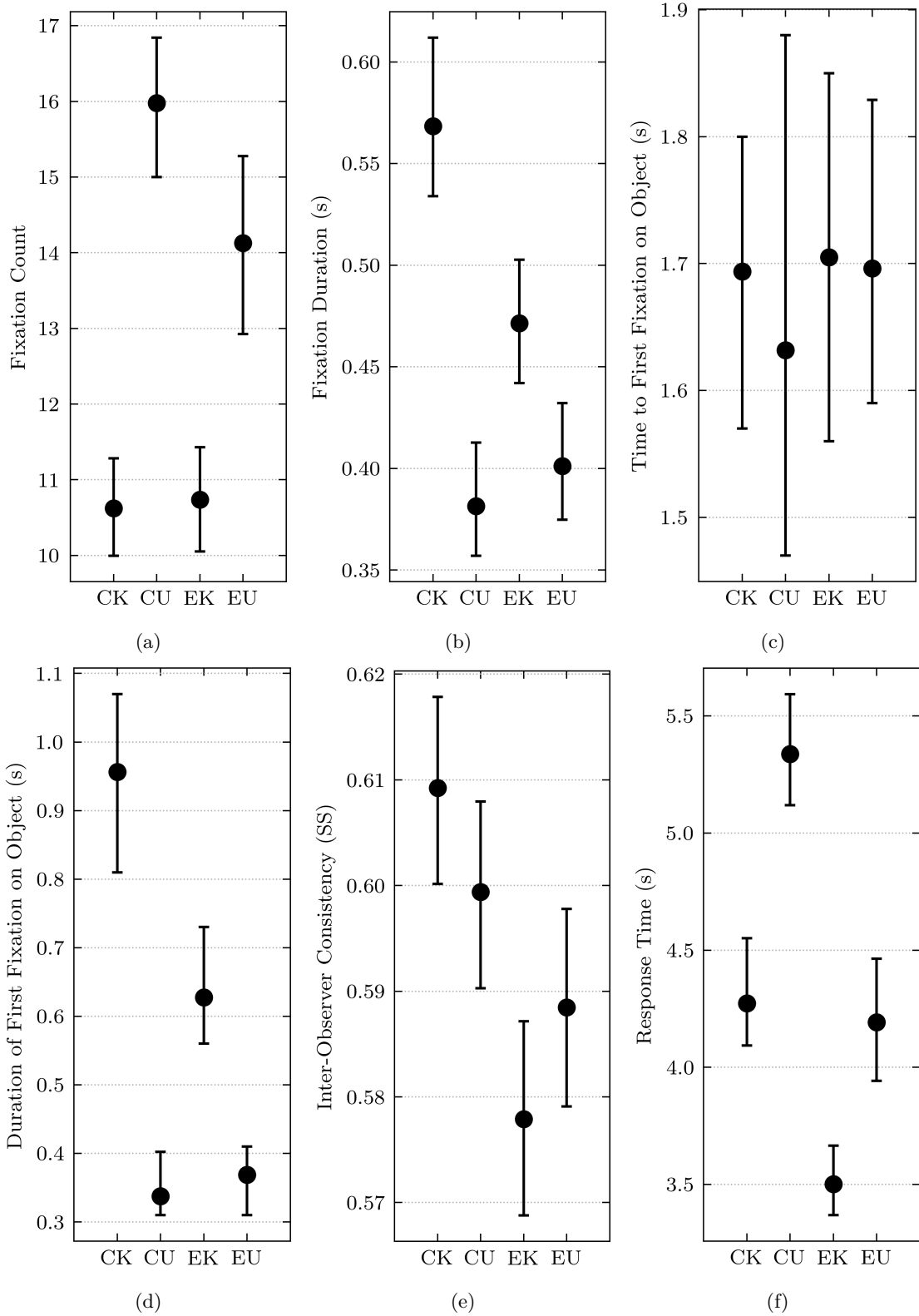


Figure 5.8: Bootstrapped means/medians and their 95% confidence intervals for the evaluated metrics for test images under the four experimental groups - exposure, no location (*EU*); exposure, location (*EK*); CCT, no location (*CU*); CCT, location (*CK*). From left: fixation counts (*Fc*), fixation durations (*Fd*), time to first fixation on target object (*TFFO*), duration of first fixation on target object (*DFFO*), inter-observer consistency using similarity score (*IOC_{SS}*)

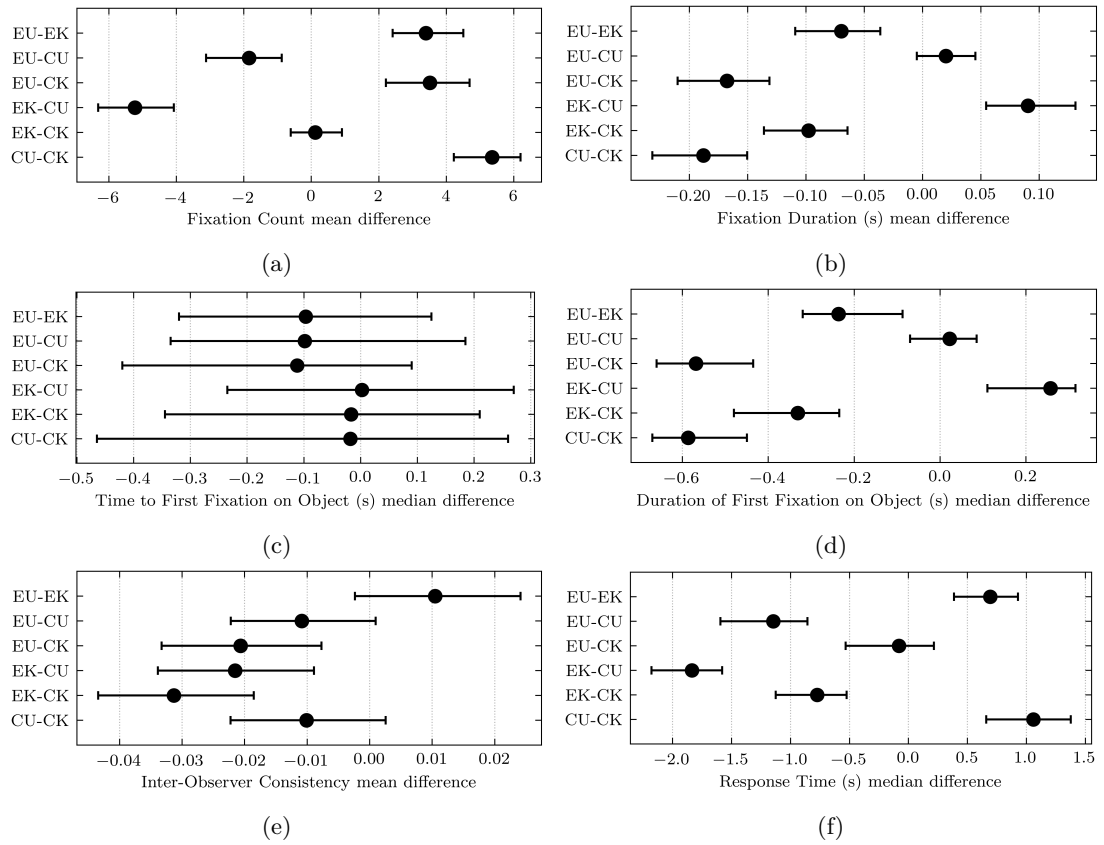


Figure 5.9: Bootstrapped comparisons of group mean/median differences for each of the evaluated metrics.

visual realism assessment and is likely induced by constraints of the task itself - finding evidence of transformations, or comparison thereof with other elements of the scene. The fact that fewer fixations are deployed when prior object knowledge is provided also indicates that visual search is relied on to identify the object, but likely not relied on during the realism rating task.

5.4.3 Fixation Durations

The average duration of fixations highlights a similar scenario to F_c , where F_d were significantly longer for conditions with prior object knowledge. This indicates that when observers are aware of the identity of the object, they deploy fewer, but longer fixations, likely due to knowledge of the existence of task-relevant information, confirming $H2$.

Specifically, for both condition EU and CU F_d were significantly shorter compared to conditions EK and CK ($p < 0.05$). Longest F_d were recorded for the CK condition, being significantly longer than all other groups ($p < 0.05$). This mirrors the behaviour observed in the case of F_c for the CU condition being significantly higher than those for the EU condition. Combined, these findings suggest that a) observers deploy more fixations searching for CCT transformations, and b) even when explicitly provided with location information, observers deploy longer fixations to composites affected by CCT transformations, as compared to exposure. No significant differences were found between conditions EU and CU . Overall fixation duration only changed significantly as a function of transformation type for groups where the object was known.

Without prior object knowledge, the average duration of deployed fixations did not change as a function of transformation type. This shows that without knowledge of the transformed object, observers deploy overall more fixations of shorter durations, suggesting dominance of visual search under this experimental condition. See Figures 5.8b and 5.9b for an illustration of these statistics and effect sizes.

5.4.4 Time to First Fixation on Object

While significantly lower $TFFO$ measures were expected for groups with prior object knowledge, this was not observed in the results, providing no positive evidence for $H3$. No significant $TFFO$ differences were noted between any of the experimental conditions. Observers required on a median of around 1.7s to attend to the transformed object, regardless of the type of transformation or prior object knowledge. See Figures 5.8c and 5.9c for an illustration of these statistics. This suggests that either a) despite additional task-relevant information, bottom-up VA processes dominate during stimulus onset focusing attention on salient features of the scene, b) observers' VA is equally attracted to the transformed object, regardless of prior knowledge c) observers do not fixate directly on the target object, but in its vicinity, resulting in overestimated $TFFO$. This is further discussed in Section 5.5.

5.4.5 Duration of First Fixation on Object

Contrary to *TFFO*, significant differences in *DFFO* were found between groups. A similar pattern to that of F_d was followed - *DFFO* were significantly longer in groups with prior object knowledge (*EK* & *CK*), confirming *H4*. Furthermore, the *DFFO* for those groups was significantly longer than the average F_d for the same group. This was not true for groups where object location was unknown *EU* and *CU*. This shows that when observers knew the identity of the transformed object, they devoted significantly more attention to it upon first fixation. See Figures 5.8d and 5.9d for an illustration of these statistics.

Reconciling this finding with the lack of differences in *TFFO* suggests that while prior object knowledge and transformation type do not change the speed at which they attract observer attention, prior object knowledge does tend to induce longer first fixations and overall longer average fixation durations. This suggests that observers prioritise the analysis of task-relevant regions, when this information is available to them. In more natural scenarios, where such additional task-relevant information is not provided, observers tend to follow a more exploratory pattern, closely resembling visual search, as indicated by significant differences in fixation counts and durations.

5.4.6 Area of Interest Similarity

The differences in F_c and F_d suggest that some significant differences between attended regions may exist across experimental conditions. Since observers deploy more fixations when they are not aware of the identity of the object, they may attend to a larger proportion of the scene, or make a larger number of comparisons between elements in the scene. Analysis of *AOIS* indicates that observers indeed rely primarily on the object regions when performing the task. A Kruskal-Wallis test showed significant differences in proportions of fixations falling on foreground objects, between reference and test conditions, for each group ($p < 0.01$). Table 5.2 details that even with prior object knowledge (in conditions *EK* and *CK*), the proportion of fixations on the object were higher in the test image, compared to the reference image.

Pairwise joint fixation map similarity measures across conditions show that changing the transformation type has a smaller effect on the *AOIS*, compared to prior object knowledge. *AOIS* between conditions *EU* – *CU* ($M = 0.81, CI = [.78, .82]$) and ($M = .82, CI = [.81, .84]$) significantly higher, compared to all other conditions (see Fig. 5.10). This indicates that the similarity of spatial VA distribution is significantly higher for groups sharing the same setting of the prior object knowledge factor, compared to any other combination of factors. This also evidences that prior object knowledge has a higher impact on spatial distribution of VA, compared to the transformed feature type, under the conditions of this study. For example, the similarity of spatial VA distributions for the same prior object knowledge setting, but different transformation features is significantly lower. These findings support *H6* that *AOIS* will be highest for groups sharing object

Experiment	Reference / Test	$F_{obj}\%$	σ	χ^2	p-value
EU	Reference	13	8	9.51	0.002
	Test	26	16		
EK	Reference	34	13	11.50	0.001
	Test	45	14		
CU	Reference	13	8	8.81	0.003
	Test	24	15		
CK	Reference	38	14	11.24	0.001
	Test	50	13		

Table 5.2: Results of comparing proportions of fixations on target object ($F_{obj}\%$) between reference and test images, under each of the experimental conditions.

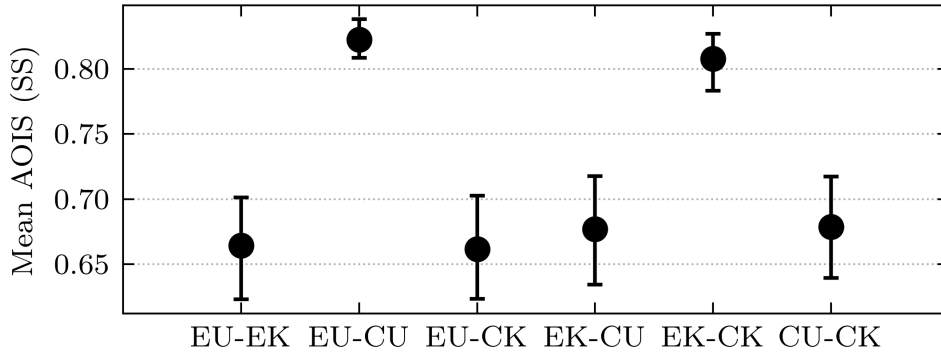


Figure 5.10: Test image area of interest similarity score (AOIS) for each pairwise combination of conditions.

knowledge priors.

Finally, qualitative analysis of attention distributions further illustrates such spatial strategies adopted by observers. During viewing of reference images without prior object knowledge (Figs. 5.11a, 5.11c, 5.12a, 5.12c), observers followed a free-viewing visual strategy, attending most objects in the scene in an exploratory manner. This was consistent across both transformation features. Conversely, when observers were provided prior object knowledge (Figs. 5.11e, 5.11g, 5.12e, 5.12g), they focused their attention on the target object, while devoting less attention to the rest of the scene. During viewing of test images, observers with prior object knowledge (Figs. 5.11f, 5.11h, 5.12f, 5.12h) focused largely on the target object, whereas observers with no prior object knowledge devoted more attention to the rest of the scene (Figs. 5.11b, 5.11d, 5.12b, 5.12d). For images with transformations of higher magnitudes (e.g. 5.12), observers' attention is often focused on the target object, even when no prior object knowledge is provided. This is understandable, since, as discussed in Chapter 4, distortions of higher magnitude tend to correlate with more accurate and confident predictions of visual realism. In cases of lower-magnitude transformations, when no prior object knowledge is available, the visual

search pattern persists as observers continue to look for evidence of a transformation. This effect is illustrated when comparing Figures 5.11b and 5.12b.

5.4.7 Inter-Observer Consistency

Overall *IOC* measures were similar across conditions (see Fig. 5.8e), however, not enough evidence is present to support *H6*. Fisher’s permutation test found significant *IOC* differences between conditions *EK* and *CK*, *EK* and *CU*, as well as *EU* and *CK* ($p < 0.05$). While, their effect sizes are small, with the largest being between conditions *EK* and *CK* ($M = -0.03$, $CI = [-.04, -.01]$), there is indication that observers were more spatially consistent with the rest of their group when assessing CCT transformations with prior object knowledge. This interpretation is in agreement with the long F_d identified for the same condition (*CK*). One explanation for this difference in *IOC* is that this additional task-relevant information reduces the within-group variance of spatial VA distributions. However, this explanation does not fit the corresponding *IOC* values for conditions *EK* and *EU*. Here, the group with prior object knowledge obtained a lower *IOC* score, compared to the group without this knowledge. See Figure 5.9e for illustration of this. While this result does not conflict with prior findings in this study, further research is required to understand the source of these significant differences.

5.4.8 Response Time

Due to the time constraint applied to the task, *RTs* can provide a good proxy for the relative cognitive load and difficulty of the tasks performed by each group. Figure 5.8f shows bootstrapped mean *RTs* and their 95% confidence intervals. It can be seen that for each transformation feature, mean *RTs* were significantly lower when object location was known. Prior object knowledge reduced mean *RT* by ~1 second. However, significant *RT* differences were also found between the transformation feature conditions. Specifically, with no prior object knowledge, observers took on average 1.15s ($CI = [0.85, 1.57]$) longer to provide a response for CCT transformations, compared to exposure transformations. When object knowledge was provided, this difference dropped to 0.77s ($CI = [0.51, 1.13]$). *RTs* for CCT transformations under each location condition were significantly higher compared to those for exposure transformations (see Fig. 5.9f). This provides further evidence that CCT was overall the more challenging feature, compared to exposure, as well as showing that prior object knowledge makes the task significantly easier to perform, for both CCT and exposure. Perhaps the most interesting aspect of this finding is its relationship with realism responses, particularly when varying the prior object knowledge factor. Gaze metrics indicate that without prior object knowledge, the task was likely more challenging than when this knowledge was present, for both exposure and CCT. However, comparing group realism ratings across the prior object knowledge condition did not reveal any significant differences. This suggests that, while gaze metrics, response times and accordingly visual strategy may be influenced by prior knowledge and deployment of VA, the resulting realism ratings do not change significantly. This and other findings are

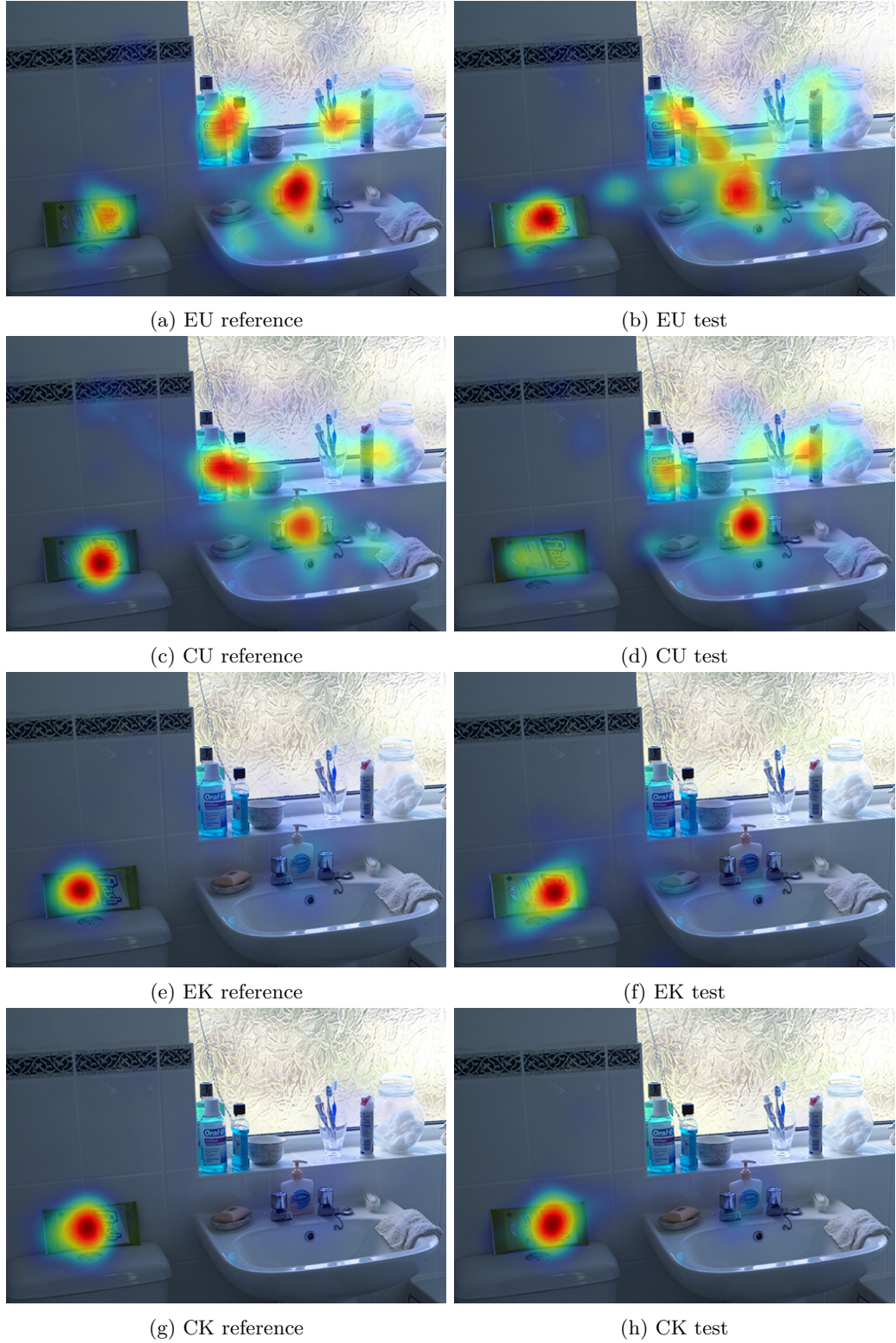


Figure 5.11: Comparison of joint fixation maps over reference and test stimuli, for each combination of factors. Test images contain the smallest positive transformation magnitude used in our experiments (+40 mired for CCT and +0.33 stops for exposure). The small transformation magnitude leads to visual search patterns in both reference and test images for conditions with no prior object knowledge (EU & CU), and increased visual effort in conditions EK and CK.

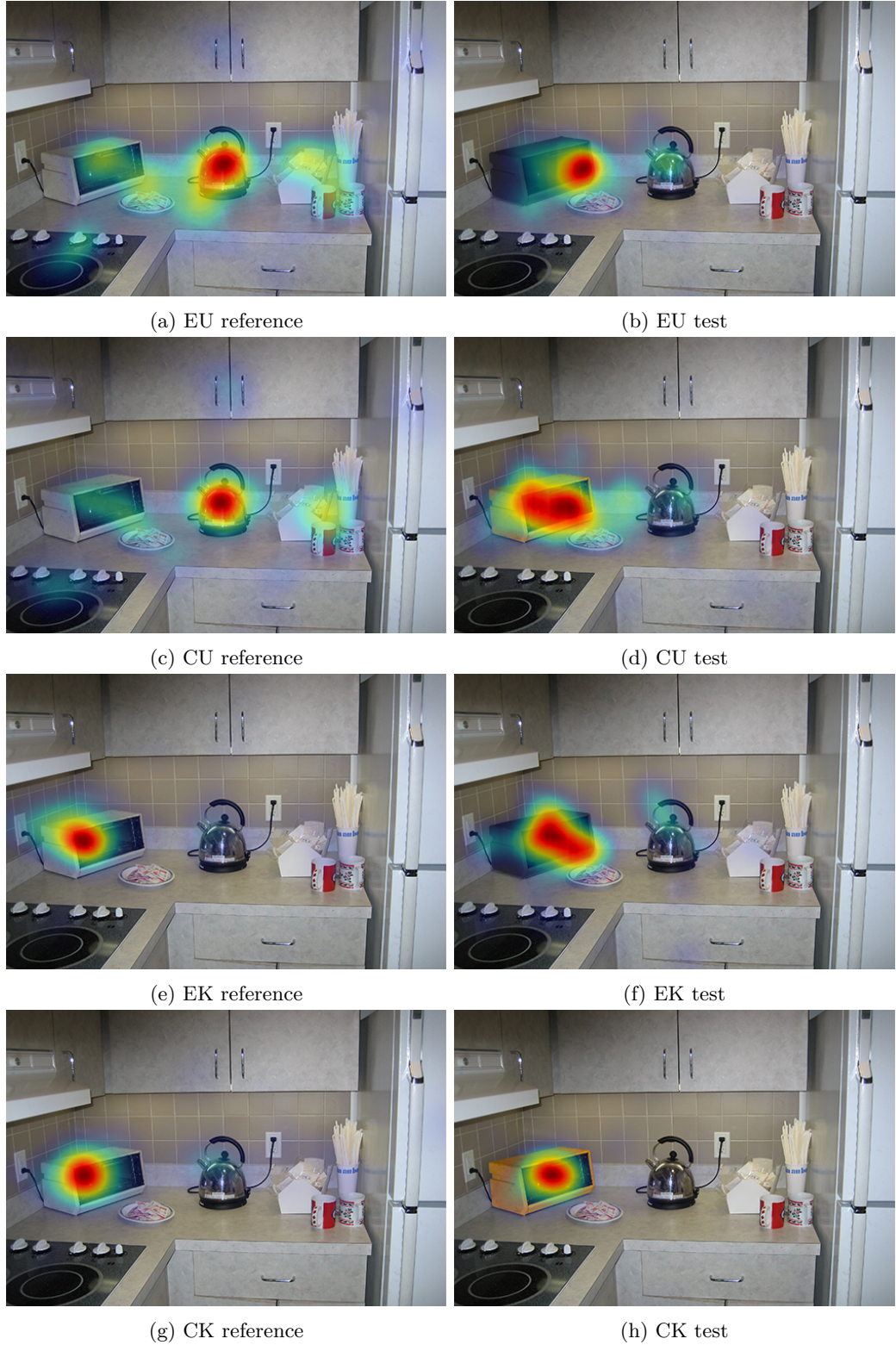


Figure 5.12: Comparison of joint fixation maps over reference and test stimuli, for each combination of factors. Test images contain the largest positive transformation magnitude used in our experiments (+200 mired for CCT and +1.66 stops for exposure). The large magnitude transformation attracts more fixations to the object, even with no prior object knowledge.

summarised and discussed in the following section.

5.5 Discussion

This section discusses the key findings of the study presented in this chapter and draws relationships with related work in the area, as well as the study from Chapter 4.

5.5.1 Summary of findings

Key findings from the presented experiments are summarised below, and discussed in depth in the following sections.

- prior object knowledge has no significant impact on subjective realism ratings
- task-relevant knowledge does have a significant impact on deployment of VA
- observer strategy can be described as a combination of visual search and discrimination
- visual search is performed to identify the object, but does not impact its realism rating
- when assessing realism, observers rely significantly on the target object, this behaviour is emphasised when the target object is known a priori.
- first fixations are significantly longer when the target object is known
- prior knowledge results in increased similarity of attention maps and shorter task completion time
- prior knowledge has a significantly higher impact on allocation of VA, compared to varying the transformed feature.

5.5.2 Prior Knowledge, VA and Realism

Chapter 4 posed several questions regarding the impact of VA on composite realism assessment and the resulting perceptual models. Chief among these was the impact of prior object knowledge on VA distribution and, consequently, on subjective realism ratings. The results presented here indicate that for the evaluated conditions, prior object knowledge did not have a significant impact on realism ratings. While significant differences in both temporal and spatial VA distribution were found between groups, realism ratings remained unchanged.

The changes in VA induced by the presence of prior knowledge, such as relatively large fixation counts and short fixation durations, can be explained primarily by the existence of a visual task, which results in VA patterns unlike those of free-viewing (Ninassi et al., 2006). Furthermore, the presence of a target (i.e. the prior object knowledge factor)

imposes even more significant top-down VA cues, essentially directing observer attention to the target object (Yarbus, 1967). The insignificant impact of prior knowledge and resulting changes in VA on realism ratings is more challenging to explain.

One explanation may be that transformations rendering a composite unrealistic are visible to observers, regardless of where their attention is initially directed, while low magnitude transformations, which do not result in unrealistic appearance of objects, may not be detected reliably, regardless of allocation of VA. It is possible to imagine that in borderline scenarios, where the transformation is very close to $1JND$, VA may have a more significant effect. However, for the images used in this study and the magnitudes of transformations evaluated, the impact of VA on realism ratings is negligible.

Another explanation is the potential impact of local transformations, or distortions, on saliency (Wang et al., 2014). For example, Leveque, Zhang and Liu (2019) showed that the saliency map of a distorted version of an image is different to that of an undistorted version of the same image. This explanation also helps explain the lack of variation in *TFFO* between groups with and without prior object knowledge in the presented experiment. If transformations applied to objects indeed affect saliency, then observer attention could be attracted to the target object before top-down task-related VA takes over. This would also explain the significantly longer *DFFO* measures for conditions with prior object knowledge: observers remain fixated on the object for longer, as they know it is the target. In the scenario where prior object knowledge is not provided, the object attracts initial bottom-up attention, but task constraints override this initial deployment of attention. These findings are in line with previous research. For example, Le Meur et al. (2010) measured overt attention of observers, both free-viewing and assessing video sequences in terms of quality. They found that while the task does influence visual attention allocation, low-level image features still play a crucial role in guiding attention, both under free-viewing and task-driven conditions.

5.5.3 Observer Strategy

The results of this study show that observers perform at least two distinct visual tasks: serial visual search, attending to individual objects in the scene, and more selective visual discrimination, once the target object is identified. As illustrated by the presented results, observers with prior knowledge of the object spent less time performing visual search and more time processing the target object, but this task dichotomy exists even when the target object is not known a priori.

Based on the feature integration theory (Treisman and Gelade, 1980), discussed earlier in this chapter, two types of visual search are performed by human observers - feature search and conjunction search. The former is based on a comparison of a single feature and is thus fast and pre-attentive, the latter relies on assessment of multiple features, is a serial process and requires allocation of overt attention. The relative involvement of these two types of

search in the context of image composite realism assessment is largely dependent on the type and magnitude of the transformation. It is not difficult to imagine an object with a drastically different appearance from its background ‘popping out’ and instantly attracting attention. Equally, one can imagine a realistic composite, where considerable effort would have to be devoted to detect inconsistencies. Thus, it is possible that depending on the severity of the transformation, or more generally, visual mismatch between the object and scene, both visual search mechanisms may be relied on.

This would provide further support to the findings discussed above, particularly the lack of significant *TFFO* differences between groups. It is logical to suggest that transformations of large magnitudes diminish the usefulness of prior object knowledge, by making the target object obvious, resulting in reduced differences between gaze metrics. On the other hand, transformations of lower magnitudes may still attract early attention by inducing changes to the saliency map, but require additional conscious effort and comparison with other elements of the scene, before a confident response can be given. This interpretation is also supported by the increased *DFFO* for conditions where object location was known.

Overall, it appears that by adopting a combination of visual search and selective longer fixations on potential targets, observers accumulate relevant visual information, based on which a final realism judgement is made. Additional attention cues, either top-down (such as prior object knowledge) or bottom-up (such as significant, attention-grabbing distortions) can accelerate this process, by reducing the contribution of visual search and overall task times.

5.5.4 Object & Scene

Based on overt VA patterns, observers seem to extract a significant amount of task-relevant information from the object and its immediate surroundings. As mentioned above, the proportion of fixations on the target object versus the rest of the scene was higher both when prior object knowledge was provided, as well as when transformations were easily noticeable (as seen in the qualitative analysis). This is also consistent with the findings of Redi et al. (2011), who showed that observer VA patterns deviate away from free-viewing patterns proportionally to the subjective quality level. As the subjective quality decreases, fixation patterns tend to deviate further from free-viewing, likely affected by changes to the saliency map. Similarly to the work presented here, they found that the distortion type has much less effect on fixation distributions, compared to the associated quality score. The interplay of scene content, attention and subjective ratings has also been investigated by Liu and Heynderickx (2011), who evaluated the impact of saliency-based weighting of objective quality scores. They found that the gain in quality prediction afforded by incorporation of saliency maps was variable as a function of image content. While further work is required to investigate the impact of scene content in detail, some trends can be gleaned from the study presented here. For example, scenes with many objects and low transformation magnitudes are likely to receive higher realism scores. This

is because the larger number of objects requiring attention under the same time constraints would likely result in shorter fixation durations, which, in turn, could lead to certain distortions remaining unnoticed and affecting the realism score positively. However, when the distortion magnitude is high, observers tend to focus their attention on the affected region.

5.5.5 Correlates of Visual Effort

Analysis of gaze metrics also reveals some information about visual effort and task difficulty. In the case CCT transformations, observers response times and fixation durations were consistently longer than for exposure transformations, when object information was revealed. This suggests that even when the target object was unambiguous, observers allocated more visual effort to rating realism of CCT transformations. This is likely related to the stimulus ranges selected for both features being unaligned. While the most extreme exposure and CCT offsets received comparable realism ratings (see 5.7), the responses for intermediate stimulus values deviate more between transformation features. In post-test conversations, particularly for condition, *CU* observers reported difficulty in spotting the transformations and assessing whether they were realistic, given the reference image. This is confirmed by the significantly higher fixation counts for that condition, indicating visual search patterns. Scene content is likely to also contribute to this effect - depending on the appearance of the object and its immediate surroundings, an exposure shift applied to that object may be more perceptible compared to a CCT shift. Redi et al. (2011) suggest that high magnitude distortions may simply require much less subjective inspection before arriving at a confident decision. This statement is reinforced by the significant fixation count, duration and response time differences between conditions. While mean fixation counts and durations between exposure and CCT transformations are not significantly different when prior object knowledge is provided, the differences becomes significant when prior object knowledge is not provided. Despite the existence of such significant differences in gaze metrics and response times, realism ratings are not affected. This suggests that given sufficient task time, observers are likely to arrive at the same realism judgements, provided they are able to detect the task-relevant signal (a suprathreshold transformation).

5.6 Conclusions

This chapter has investigated the deployment of visual attention during subjective realism rating tasks in the context of local transformations approximating common image composite distortions. To achieve this, eye movement, task performance metrics and subjective realism ratings were collected for observers performing a subjective realism assessment task. The task required observers to provide a binary realism rating for images affected by local transformations. It measured the impact of two factors: prior object knowledge and local transformation type on the resulting realism ratings and eye

movement metrics.

The findings of this study indicate that when rating the realism of images affected by local exposure and CCT transformations, observers rely heavily on the features of the object affected by the transformation and its immediate surroundings. Despite the importance of this area to the task, its explicit prior indication does not have a significant effect on the resulting realism ratings. It does, however, impact response times and fixation metrics associated with visual effort, task-relevance and observer confidence. While realism ratings across both transformation features were moderately correlated, there is indication that, in the stimulus ranges under evaluation, CCT transformations required more visual effort for observers to detect and assess, compared to exposure transformations. Despite the fact that some of these differences (e.g. fixation counts) diminish when the target object is made explicit to observers, significant differences in response times persist, suggesting that a simple distortion-magnitude-based model may not generalise well across multiple local transformation or distortion types and different scenes. Due to a visual-search-based strategy being deployed in the context of low-magnitude transformations, realism ratings may be positively-biased for visually busy scenes, containing many objects for observers to visually compare and contrast, given a time constraint.

Given the negligible impact of prior knowledge and VA deployment on realism ratings, the realism rating for an image affected by a local composite-like transformation seems to largely be a function of inter-observer variability, scene content and visible object-scene appearance differences. As indicated in Chapter 4, the visibility of such differences can be modelled based on type and magnitude of the transformation (or distortion), leading to the conclusion that for a given scene, distortion type and magnitude, realism can be modelled in terms of group JNDs for that combination of object, scene and distortion.

Chapter 6 will investigate whether the findings from Chapters 4 and 5 can be generalised into a perceptual model capable of approximating human sensitivity in detecting local transformations in images.

Chapter 6

Modelling Image-Wise Observer Sensitivity

This work was published in:

Dolhasz, A., Harvey, C. and Williams, I., 2020. Learning to Observe: Approximating Human Perceptual Thresholds for Detection of Suprathreshold Image Transformations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp.4797–4807

6.1 Introduction

Chapters 4 and 5 have illustrated how subjective realism can be modelled as a function of just-noticeable distortions (represented by local transformations), and which visual information observers rely on in order to produce those subjective realism judgments. While the presented studies provide insight into observer sensitivity and strategy, they rely on relative transformations to create composite-like distortions from real photographs. In order to apply the JND model from Chapter 4 to a real composite image, the transformation magnitude must be known. This requires access to the *ideal* appearance (the *original* version of the image in our experiments). Clearly, this is infeasible to acquire in the case of a real composite image. Thus, in order to make the proposed models applicable and generalisable to real composites, they must not rely on knowing the ideal appearance of the composite a priori.

Accordingly, this chapter revisits and adapts the JND models from Chapter 4, based on findings from Chapter 5. Through modelling group JND for *individual images*, as a function of local transformation magnitude, the impact of scene content and underlying object appearance is encoded directly in the JNDs. This then allows the representation of the group realism rating process as a nonlinear function of the input image content, object appearance and local transformation type and magnitude. This nonlinear function

can then be approximated using convolutional neural networks, in order to generalise performance to new, unseen images. This model is then evaluated with respect to empirical JNDs and applied to real image composites. This approach to modelling detection of image composite regions requiring harmonisation also relaxes the requirements for input masks. This allows to address the limitations of existing methods, discussed in Section 3.5, by replacing the input mask with one that can be detected during inference, using the model developed in this chapter. This, combined with state-of-the-art harmonisation techniques, allows for new applications for image content for which input masks are not available, such as legacy photographs and films. Commonly, such content would otherwise require human intervention to manually create these masks.

6.2 Related Work

6.2.1 Learning-based Human Perception Models

Many complex visual tasks, such as object detection or face recognition, are performed effortlessly by humans, but have long been difficult problems in computer vision. Depending on perspective, illumination and scene context, the projection of a given three-dimensional object onto a two-dimensional retina (or camera sensor) can take on a range of different appearances. Likewise, a given 2-D retinal projection can be the result of infinitely many arrangements of 3-D objects in the original scene. This in turn makes such problems near impossible to solve using formal, rule-based methods, because of the sheer number of rules that would have to be designed. In addition to this, collection of perceptual data is commonly a costly and time-consuming process, as it relies on use of human observers. Generalisable perceptual models thus have the capacity to remove the requirement for humans in the loop, making them attractive, both for practical applications as well as research into human perception.

In recent years, the rapid development of machine learning has contributed to solving many such problems by approximating the function mapping input stimuli (e.g. images), or certain features thereof, to corresponding labels (e.g. object classes), based on exemplar data. In computer vision and image processing, much of this progress has been achieved using deep convolutional neural networks (see Section 3.3.5 for a review). The advantage of learning-based methods in this context is that they allow modelling of complex processes using exemplar data, without the need for explicit programming. For example, Bosse et al. (2017) trained CNNs to perform both full-reference and no-reference image quality analysis, based on datasets of images and associated human opinions (Sheikh et al., 2005; Ponomarenko et al., 2013). Talebi and Milanfar (2018) extend this approach, by mapping images to distributions of opinion scores, in order to account for subjective variabilities. Similar approaches have been applied to other subjective qualities, such as image aesthetics (Kong et al., 2016; Ma, Liu and Wen Chen, 2017) or perceptual image similarity (Zhang et al., 2018b). Provided that a large dataset of training data and associated labels exist,

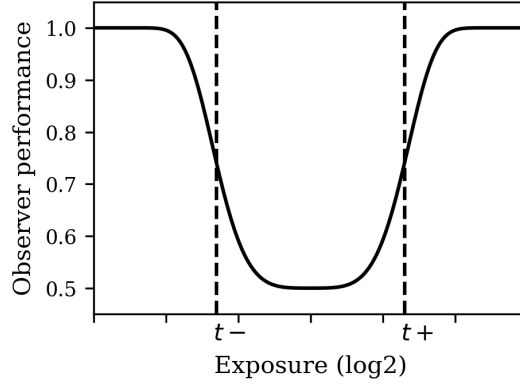


Figure 6.1: Illustration of just-noticeable differences for negative ($t-$) and positive ($t+$) local exposure transformations as a function of exposure shift. Exposure shifts of magnitudes between $t-$ and $t+$ are perceptually subthreshold, while magnitudes outside that range are perceptually suprathreshold. These three ranges are used to define the class boundaries in terms of local transformation magnitude.

DNNs perform very well at approximating the functions mapping the data to labels.

While supervised deep learning algorithms can achieve high accuracy when learning to map images to subjective opinion scores, they require large amounts of labelled training data. As discussed in previous chapters, the collection of subjective perceptual data under well-controlled conditions is a time- and effort-consuming process. Many approaches have been proposed to tackle the problem of small datasets, including data augmentation (Perez and Wang, 2017), transfer learning (Pan and Yang, 2009) and semi-supervised methods (Zhu, 2005).

6.2.2 Unsupervised, Self-supervised & Transfer Learning

As discussed in Section 3.2.4, unsupervised learning algorithms relax the requirement for labelled training data. In the context of images, their common applications include representation learning, the goal of which is to encode images into compact feature representations useful for related tasks. This has been used extensively in image synthesis methods, for example, to drive a generative adversarial network (Radford, Metz and Chintala, 2015), or to improve classification performance in mammography images (Arevalo et al., 2016). Self-supervised approaches have recently proven practical for this task. In self-supervised learning, the training process follows a conventional supervised paradigm, however the training data is generated and/or labelled automatically. In the case of images, this commonly involves applying some known transformation to the training data and training the network to predict this transformation, or encode the same features despite the transformation. For example, Noroozi and Favaro (2016) generate jigsaw puzzles from images and learn good representation through solving these, Doersch, Gupta and Efros (2015) predict context from a set of candidates, whereas Gidaris, Singh and Komodakis (2018) predict image rotations.

These methods exploit the assumption that a compact visual feature representation trained to counteract or predict certain input data transformations on a large enough dataset of unlabelled images is likely to perform well at related visual tasks, such as image classification, or captioning. This is also the conceptual foundation for transfer learning, where a model trained for a particular task on a large dataset is re-trained for a related task, on a different dataset (Pan and Yang, 2009; Weiss, Khoshgoftaar and Wang, 2016). Transfer learning thus exploits a key property of hierarchical feature representations: while high-level features are specific and vary significantly between various tasks, low-level features are more general and tend to be similar across tasks. In practice, transfer learning often involves *freezing* of the weights of lower-level layers of a pre-trained model and fine-tuning only the last few layers. This way the general low-level features are preserved, while the high level features are adapted to the domain of the new task. It has also been shown that the activations of DNNs trained on object recognition tasks correlate with inferior temporal cortex spiking responses to natural images (Yamins et al., 2014), suggesting that models trained on visual tasks, tend to respond to similar features as the neurons in the visual cortex.

6.2.3 Semi-supervised Learning

Semi-supervised learning, combines the advantages of both supervised and unsupervised learning. This is commonly accomplished by learning a representation in an unsupervised manner and training a supervised model using the learned representation and a small dataset of labelled examples. Semi-supervised methods which impose further constraints on the number of labelled training samples are commonly referred to as *few-shot learning* (Sung et al., 2018), *one-shot learning* (Vinyals et al., 2016) or *zero-shot learning* (Socher et al., 2013), depending on the number of labelled training samples presented to the model during training. These methods commonly rely on first learning a good task-relevant representation of input data, followed by performing an association between a particular region of the representation feature space with a particular class label (Snell, Swersky and Zemel, 2017).

6.2.4 Invariance & Equivariance

Depending on the task at hand, the learned representation should be *invariant* to certain aspects of the input, while being *equivariant* to other properties. For example, an image classification CNN should be approximately translation invariant - regardless of the spatial position of a cat in an input image, the CNN's output should always indicate a cat. This is in spite of the fact that the building block of a CNN – the convolution operation – is a translation equivariant operation, in that the translation of the input results in an equal translation of the output. CNNs for image classification achieve translational invariance through the use of max pooling layers (Boureau, Ponce and LeCun, 2010). However, invariance to other properties of the input is not always developed through architectural modifications. Data augmentation is a common approach to introducing

invariance to certain input properties during training of the model. For example, in order to make an image classifier invariant to the rotation of the input, it is common to generate multiple variants of input samples at random orientations, but with the same training labels (Schmidt and Roth, 2012).

The properties of invariance and equivariance to certain transformations of the input are not universally desired in every application. As opposed to image classification, in semantic segmentation, the neural network should be equivariant to translations of the input, since the output class map must align with the locations of the semantic regions in the input image (Long, Shelhamer and Darrell, 2015). Analogously, a network tasked with predicting a subjective response given some transformation an input image, should be equivariant to that input transformation.

6.2.5 Transformation Equivariant Representations

In addition to the architecture- and augmentation-based approaches, several generalised approaches to transformation-equivariant representations (TERs) learning (Hinton, Krizhevsky and Wang, 2011) have been proposed. Specifically, Zhang et al. (2019); Qi et al. (2019) show that for image-to-image problems, robust transformation-equivariant representations (TERs) can be learned by auto-encoding transformation parameters, as opposed to auto-encoding data augmented by input transformations. They show that representations equivariant to both parametric and non-parametric transformations can be learned using conventional convolutional autoencoders with minor changes to the training protocol. Specifically, they use a Siamese architecture (Bromley et al., 1994; Chopra, Hadsell and LeCun, 2005) consisting of a convolutional encoder. This is then trained to regress the parameters $\hat{\theta}$ of a projective transformation $T(\theta) \in \mathbb{R}^{3 \times 3}$ between an input image I and the transformed image $\hat{I} = T(I; \theta)$. Since the parameters of the transformation are generated at run-time, they can be used as dynamic ground truth labels, making the process self-supervised. Aside from its generalisability, the authors demonstrate superior results on auxiliary tasks, such as image classification, compared to alternative methods.

As illustrated in Chapter 4, many distortions present in image composites can be modelled in a controlled manner by applying local transformations to real images. Similarly, Chapter 2 discussed how compositing artists commonly improve the realism of image composites through manual application of local transformations. Accordingly, feature representations of images used in computational models of realism should be equivariant with respect to such transformations. This, in turn, allows for subsequent perceptual weighting of such a feature space to align with human perceptual thresholds. In other words, for a model to reliably detect unrealistic image composites, it must be capable of first encoding the presence and magnitude of such transformations. However, no previous work has leveraged this observation for modelling of subjective image properties.

Accordingly, this chapter develops a methodology to learn representations equivariant to local transformations, by adopting the paradigm of autoencoding transformations (Zhang et al., 2019). By leveraging DCNN-based techniques and synthetic data generation, these TERs can be learned in a self-supervised manner for a large dataset of images. This TER can then be leveraged to learn a mapping between image composites and corresponding image-wise JNDs.

6.3 Method

6.3.1 Overview

This section details the methodology for learning a mapping between input images affected by local exposure transformations and perceptually-based transformation maps, indicating the pixel-wise probability of the existence of suprathreshold local image transformations. Building on the methodology and findings from Chapters 4 and 5, detection of local image transformations is modelled as a pixel-wise classification problem, where each pixel of the input image is assigned one of three labels: negative suprathreshold, positive suprathreshold and subthreshold. These categories are based on image-wise perceptual thresholds, which define the decision boundaries of the model (as illustrated in Figure 6.1). By estimating psychometric functions and JNDs in an image-wise fashion, the impact of the object, scene and transformation type are encoded in the subjective response. This in turn allows the model to learn the impact of image content and local context on observer sensitivity to local image transformations in a supervised manner. Unsupervised pre-training and transfer learning techniques are also adopted in order to maximise generalisability of the proposed model, given the limited size of the training dataset.

The following sections describe the rationale behind the proposed approach and adopted methodology, including an overview of the 2AFC study and resulting psychometric functions, formulation and generation of training data based on the perceptual models, neural network architecture design, optimization details and techniques adopted to optimise performance.

6.3.2 Distortions as Transformations

Chapter 4 illustrated that many distortions affecting image quality, or realism, can be seen as transformations applied to the original, uncorrupted image as a side effect of some processes such as transmission, compositing, or compression. If the magnitude of this transformation is known, the corresponding JND model can be used to predict subjective realism ratings. However, without access to an ideal reference image, estimation of the magnitude of a local transformation is non-trivial, as its appearance is conditioned on the underlying image pixels. As discussed previously, DL-based approaches are a good fit for such problems, however, any representation mapping images affected by local transformations $I_T = T(I; \theta)$ to a point on a perceptual scale, must be equivariant to

the transformation of interest present in the input data, such that:

$$\phi(T(I)) = T(\phi(I)) \quad (6.1)$$

where I is an input image, T is a local image transformation $T : I \mapsto I$, and ϕ is a function mapping from input images to a feature representation. In other words, in order for the learned representation to be equivariant to local transformation of the input, the learned representation should be able to encode the effects of this transformation.

Perceptually suprathreshold transformations constitute a subset of a family of local image transformations, for example local exposure shifts,

$$\mathbf{T} = \{T_\theta \mid \theta \sim \Theta\}$$

parametrised by θ sampled from a distribution Θ , conditioned for a specific observer and object-scene combination. As such, detection of such transformations can be modelled in a two-stage fashion:

1. mapping of images to points in TER feature space
2. classification of points in this feature space as sub- or suprathreshold

This formulation of the problem highlights the independent nature of the representation learning task, which can be addressed using unsupervised learning methods, from the perceptually-based classification task, for which training data is limited and costly to acquire.

6.3.3 Perceptual Thresholds as Decision Boundaries

In the context of image distortions, and assuming controlled viewing conditions, the sensitivity of a given observer to different magnitudes of a local transformation can be modelled using a psychometric function, as in Eq. 4.9. This function can be seen as the result of an observer process operating on a range of input data. Given an unprocessed image I , object mask M , observer function O and \tilde{I}_x a corrupted version of I resulting from a local transformation $T(I, M, x)$, the empirical psychometric function can be interpreted as a result of applying the observer function to \tilde{I} for all values of x . The observer function O thus represents the perceptual process performed by an observer (or group of observers), which maps an input stimulus \tilde{I}_x to a point on the psychometric function.

$$\Psi(x) = O\left(T(I, M, x)\right) \quad (6.2)$$

where the observer function can also be interpreted as a combination of a nonlinear function ω mapping images to a TER, and another function ρ mapping these features

to a probability:

$$O(I) = \rho(\phi(I)) \quad (6.3)$$

Accordingly, detection of *local* suprathreshold transformations in an image can be defined as applying the observer model to classify each pixel based on the existence of the effects of a suprathreshold transformation. Under this arrangement, the decision boundaries of the classifier ρ are conditioned to approximate the thresholds x_{t-} and x_{t+} of a psychometric function Ψ , estimated with respect to the magnitude x of transformation T applied to an image region in I defined by mask M . This results in 3 classes c , defined by the parameter of the transformation generating the input stimulus:

$$c = \begin{cases} 0, & \text{if } x < x_{t-} \\ 1, & \text{if } x > x_{t+} \\ 2, & \text{otherwise} \end{cases} \quad (6.4)$$

Here, x_t is the value of the transformation parameter for which the probability of detection exceeds threshold t , set to 0.75, corresponding to the JND in 2AFC tasks. This is the midpoint between perfect (100%) and chance (50% for 2AFC task) performance Wichmann and Hill (2001b). As two psychometric functions are estimated per image, one corresponding to decreasing the pixel intensity (x_{t-}) and one for increasing it (x_{t+}), their two thresholds separate the parameter space x into three regions (Fig. 6.2d).

6.3.4 Psychometric Function Estimation

To estimate image-wise empirical psychometric functions with respect to a local exposure transformations, a 2AFC study is designed, following and extending the methodology described in Chapter 4. Specifically, for each image, the performance of observers completing an image discrimination task is measured. The task involves correct discrimination between an *original* version of the image I , and a *transformed* version \tilde{I}_x , affected by a local exposure transformation, where x specifies the magnitude of the transformation. This is performed for a range of transformation magnitudes, for each image, across multiple observers. The original (I) and transformed (\tilde{I}) images are displayed side by side in random order, and observers are requested to identify I correctly. Finally, Weibull psychometric functions are fit to each observer's responses for each image. To extract the thresholds, the parameter values x_{t-} and x_{t+} corresponding to a performance level of y_t for negative and positive exposure shifts, respectively, are estimated. Mean thresholds are then bootstrapped across all observers who viewed the same image. This process is discussed in detail in the following sections.

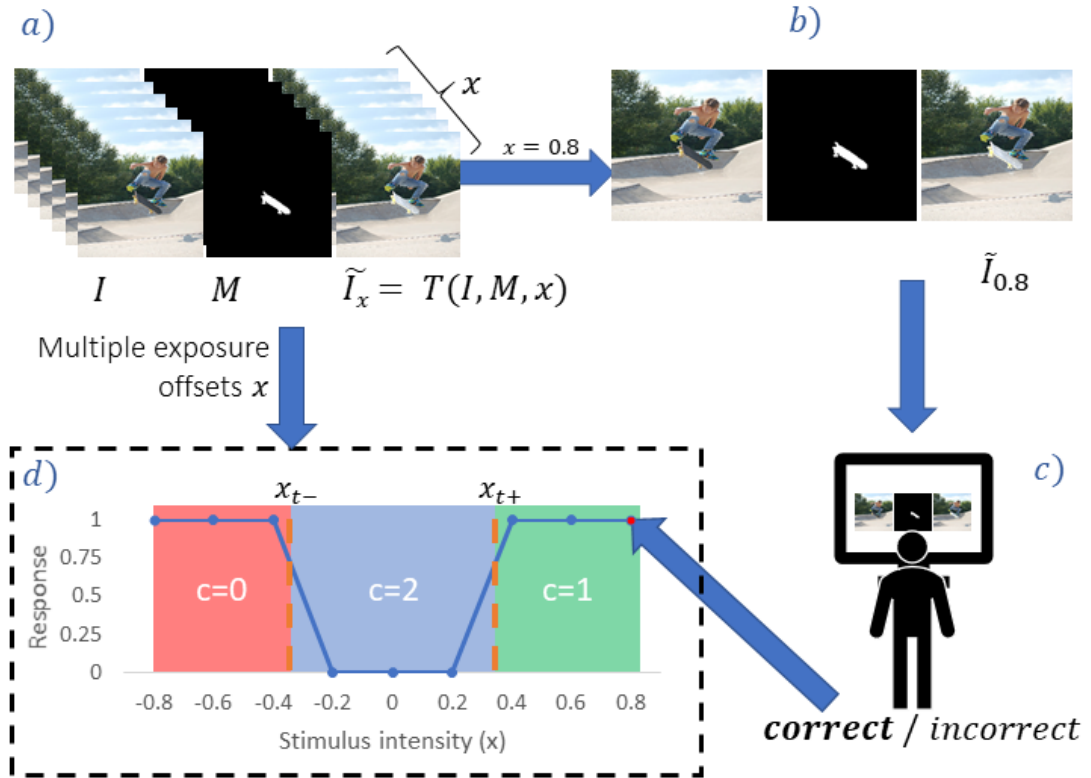


Figure 6.2: Illustration of the 2AFC procedure used in the experiments. a) For a given image I and object mask M we generate images \tilde{I} with different exposure offsets based on the sampled value of x . b) Example stimulus displayed to an observer. c) Observer correctly identifies I and \tilde{I} for $x = 0.8$. d) Observer response added to their previous responses for different sampled values of x . Symbols x_{t-} and x_{t+} , illustrated with orange dashed lines, indicate the location of the threshold after performing psychometric function fitting.

Observers

A total of $N=120$ (44 females) naive observers, with a mean age of 31 ($SD = 11.85$), are recruited from a population of University staff and students and were randomly assigned to 20 groups. Of the 120 observers, 50 declare themselves as laypersons with respect to digital image assessment, while the remaining 70 report relevant experience in graphic design, computer games or digital image processing. All observers are screened for normal visual acuity and colour vision before participating in the experiment (Zeiss, 2014).

Experiment Design

All experiments are performed under controlled laboratory conditions, following the ITU BT-500 recommendation ITU-R BT (2002). We use an Apple Cinema HD 23" monitor, calibrated to sRGB colorspace using an X-Rite i1Display Pro display calibration device. Observers are positioned 65cm away from the display. To mitigate the confounding impact of visual search on the task, particularly when differences between the images are minimal, we explicitly indicate the transformed region in the image by displaying the binary mask corresponding to the object, following the strategy from Chapter 4. To minimize the number of experimental trials, we leverage the QUEST adaptive sampling procedure (Watson and Pelli, 1983), using the implementation from the PsychoPy 2 library (Peirce et al., 2019).

Stimuli & Experimental Procedure

The stimuli dataset consists of 300 8-bit images with corresponding object masks. These include the 165 images used in Chapters 4 and 5, as well as an additional 135 images sampled from the LabelMe (Russell et al., 2008) and SUN (Xiao et al., 2010) datasets using the same strategy adopted in Chapter 4. In order to minimise the duration of the experiment for any single participant, the dataset was split into 20 batches of 15 images. These images are then evenly and randomly distributed across 20 groups of 6 observers. Each group views 15 unique images from the dataset. In the experimental session, each observer performs repeated 2AFC trials for each of the 15 base images in their allocated image sample, viewing at least 20 different variations of each base image. Observers first complete 20 trials using a calibrating image, results for which are discarded.

In each trial, observers are shown 2 images: the original image I and a transformed version of the original image \tilde{I}_x , the result of exposure transformation $T(I, M, x)$ of magnitude x . A segmentation mask M is also displayed indicating the target object. These images are displayed at the same time and remain on-screen for 5 seconds. The order of I and \tilde{I} is randomized every trial. Observers are instructed to correctly indicate I by clicking a corresponding button. After each response, a new value of x is sampled by the QUEST procedure (Watson and Pelli, 1983), and the process is repeated.

Perceptual Threshold Estimation

Binary responses y with corresponding stimulus intensities x are collected for each observer-image combination. The PsychoPy library Peirce et al. (2019) is then used to fit two Weibull cumulative distribution functions to each set of responses, one for negative and one for positive exposure shifts. This function is given by

$$y = 1 - (1 - \gamma)e^{-\left(\frac{kx}{t}\right)^\beta} \quad (6.5)$$

and

$$k = -\log\left(\frac{1 - \alpha}{1 - \gamma}\right)^{\frac{1}{\beta}} \quad (6.6)$$

where x is the stimulus intensity, y is the proportion of correct responses, γ is the performance level expected at chance, equal to 0.5 for 2AFC tasks, α is the performance level defining the threshold (set to 0.75, corresponding to the JND for 2AFC), β is the slope of the function and t is the threshold. The threshold values are pooled across observers for each image and bootstrap resampling is used to estimate mean image thresholds, using 1000 bootstrap samples. Two generalized perceptual thresholds, x_{t-} and x_{t+} , are thus estimated for each image in the dataset.

6.3.5 Transformation Equivariant Representation Learning

Many CNN models commonly used as a starting point for transfer learning, such as image classifiers trained on ImageNet, are conditioned to be invariant to changes in appearance-based object properties, such as brightness, illumination or colour. While this is a desired property for object classifiers, the task of local exposure transformation detection explicitly leverages such appearance-based features to assign classes to output pixels. Thus, transfer learning using a representation trained on an image classification task, and thus invariant to the transformation of interest, is likely to produce suboptimal results.

To address this, a task-specific TER is first learned in an unsupervised manner, adopting the auto-encoding transformations (AET) approach of Zhang et al. (2019), who encode a TER by learning to predict transformation parameters that describe a transformation between two inputs. Using this approach, a representation invariant to a particular transformation type - local exposure shifts - can be trained.

AET: Network Architecture

In order to design a NN architecture capable of learning a representation equivariant to local exposure transformations, the AET method described by Zhang et al. (2019) is first extended from prediction of global transformations to prediction of local transformations. This is achieved by changing the dimensions of the output of the network from a 3×3 projection matrix, to an $H \times W$ matrix, where each entry describes the predicted magnitude of a local exposure transformation between the two input images. Furthermore, the shallow

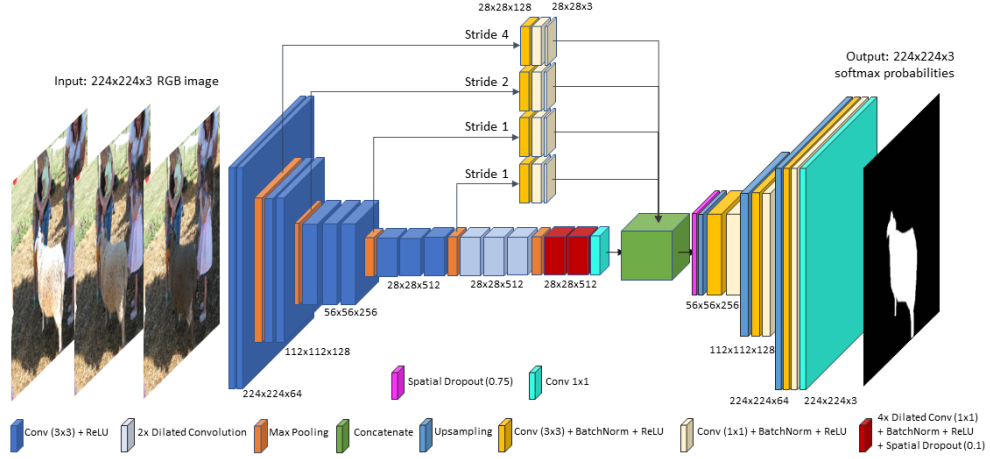


Figure 6.3: The architecture of the VGG16-based convolutional autoencoder used in the TER and perceptual threshold learning task. The network is based on a FCN adaptation of the VGG16. See Section 6.3.5 for a detailed description of the architecture.

encoder proposed by the authors is replaced by a VGG16 architecture, initialised with ImageNet weights. In order to adapt the model to an image-to-image task, the VGG16 is converted to a fully convolutional network, following Long, Shelhamer and Darrell (2015) and the success of their method in semantic segmentation tasks. Due to the importance of contextual and multiscale information, a multiscale extension is incorporated, as proposed in Li and Yu (2018). This introduces skip connections to the model, taking outputs after each max pooling layer in the VGG16 backbone and passing each through an additional convolutional branch before concatenating the output of all branches. Each branch consists of 3 convolutional blocks. The first block contains a 3×3 , 128-channel convolutional layer with a stride setting dependent on the scale of the input. This is 4, 2, 1, 1 respectively for inputs from the first 4 max pooling layers, causing all multiscale branches to output feature maps of equal resolution. Each of these layers is followed by a batch normalisation layer and a ReLU activation. The following two blocks contain 1×1 convolutional layers with a stride of 1, with 128 and 3 channels, respectively. They are each followed by batch normalization and a ReLU activation. To output masks of equal resolution to the input images, a convolutional decoder is attached to the output of the multiscale concatenation layer. It consists of 3 blocks, each block containing a $2 \times$ upsampling layer, followed by two sets of convolution, batch normalization, and ReLU layers. The first convolution in the block uses 3×3 kernels, while the second uses 1×1 kernels. See Figure 6.3 for a detailed overview.

Using this architecture, the AET model is designed by sharing the weights of the network between two image inputs, I and \tilde{I}_x (Fig. 6.4). Activations for both inputs are then concatenated and fed to a final convolutional layer. As the transformation can be expressed by a single scalar the final layer of our AET is a 3×3 convolutional layer with a linear activation, which outputs masks with resolution equal to the input image, with a single value expressing the predicted exposure shift for each pixel. This way, the model can be

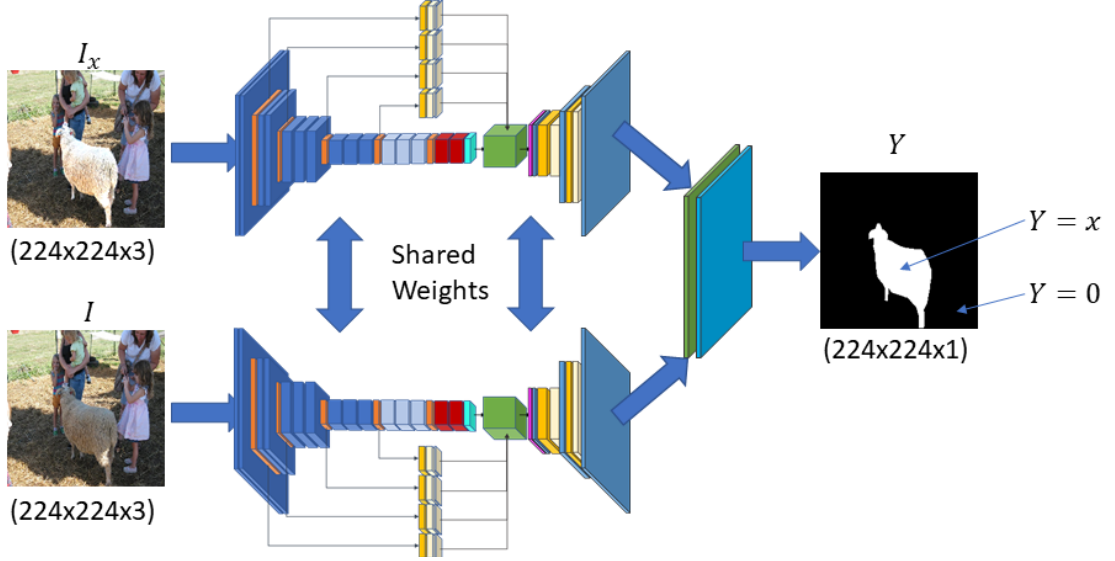


Figure 6.4: Unsupervised AET architecture consisting of a VGG16-based convolutional autoencoder with weights shared across two inputs. Activations for both inputs are then concatenated and fed to a final convolutional layer with a single channel output. The output masks encode the parameter of the transformation for each pixel.

trained to approximate pixel-wise transformations applied to an input image.

AET: Training Data Generation

To train the AET in an unsupervised manner, training data are dynamically generated at runtime. Specifically, input images I are used to generate transformed images \tilde{I} and corresponding output masks $Y = xM$, which encode the parameter of the transformation applied to the input. \tilde{I} contains an exposure shift applied within the region defined by M . Each pixel in Y contains the value of the exposure shift x applied to the corresponding pixel in \tilde{I} . This is x wherever $M = 1$ and 0 elsewhere (Fig. 6.4). During training, images I and corresponding masks M are dynamically sampled from the MSCOCO dataset Lin et al. (2014). As some images in MSCOCO contain multiple masks, one is randomly selected, provided its area is larger than 1% of the image, while the other masks are ignored. Local exposure shifts are then applied by sampling the transformation parameter x and scaling the luminance channel of I after conversion to Lab colourspace:

$$\tilde{I}_L = 2^x I_L \odot M + I_L \odot (1 - M) \quad (6.7)$$

where x is a scalar sampled from a base-2 log-uniform distribution spanning $(\log_2(0.1), \log_2(10))$, I_L is the luminance channel of the original image I after conversion from RGB to Lab colourspace, M is the alpha mask and \odot is the Hadamard product. The pixel values of the processed image are clipped to the range $(0.0, 1.0)$, converted back to RGB , rescaled to 0.0 mean and unit variance. After resizing to $(224, 224, 3)$, both I and \tilde{I} are fed to the two inputs of the AET (as in Fig. 6.4). The output of the network is a mask \hat{Y} approximating the parameter of the transformation at each pixel of the input image.

AET: Objective & Optimizer Details

The AET model is trained using the Adam optimizer (Kingma and Ba, 2014). Default values are used for all parameters, aside from the learning rate, which is controlled using a cosine annealing schedule (Loshchilov and Hutter, 2016). The minimum and maximum learning rate in the annealing schedule are set to $1e-6$ and $1e-4$, respectively. The learning rate cycles between these values over 5 epochs, after which the maximum learning rate is reduced to 90% of its value, and the cycle is repeated for $1.5\times$ as many epochs. Overall, the AET is trained for 90 epochs, minimizing the mean squared error (MSE) loss (as defined in Eq. 7.1) between \hat{Y} and Y . The model with the lowest validation error is then selected for use as the backbone for the Perceptual Threshold Classifier.

6.3.6 Perceptual Threshold Classifier (PTC)

The Perceptual Threshold Classifier (PTC) corresponds to the function ρ from Eq. 6.3. Given a TER of the input image, encoded using the AET (ϕ in Eq. 6.3), the PTC assigns a class to each input pixel, based on whether it is affected by a suprathreshold transformation.

PTC: Network Architecture

Given the trained AET model described in Section 6.3.5, the encoder and decoder shown in Figure 6.3 are extracted and the final single-channel convolutional layer of the decoder is replaced with a spatial dropout layer with a dropout probability of 75%, followed by a 3-channel convolutional layer with a softmax activation. The AET-specific elements, such as the second image input and the feature pooling and concatenation layers, are removed. As such, the network resembles a conventional convolutional autoencoder. The relationship between the architecture of the AET and the PTC is illustrated in Figure 6.5.

PTC: Training Data Generation

Using the image-wise perceptual thresholds collected in the 2AFC experiment (see 6.3.4), a data generation method is devised which dynamically applies random exposure transformations to the images used in our 2AFC experiment and generates corresponding class maps, based on whether the parameter of the transformation x exceeds one of the two empirical thresholds estimated for a given image. When x exceeds a threshold, any pixels affected by this suprathreshold transformation are assigned $c = 0$ (negative suprathreshold exposure shift) or $c = 1$ (positive suprathreshold exposure shift), following Equation 6.4. The last channel of the target image corresponding to $c = 2$ is conceptually similar to the *background* class in semantic segmentation models, indicating pixels that do not belong to any of the foreground classes. In our case, these are pixels unaffected by a suprathreshold transformation. We use a 90%-10% training/validation split. The shape of the target mask is $(224, 224, 3)$, containing one channel per class. During training, a data generator constrained to ensure a balanced class distribution in each minibatch is used. Specifically,

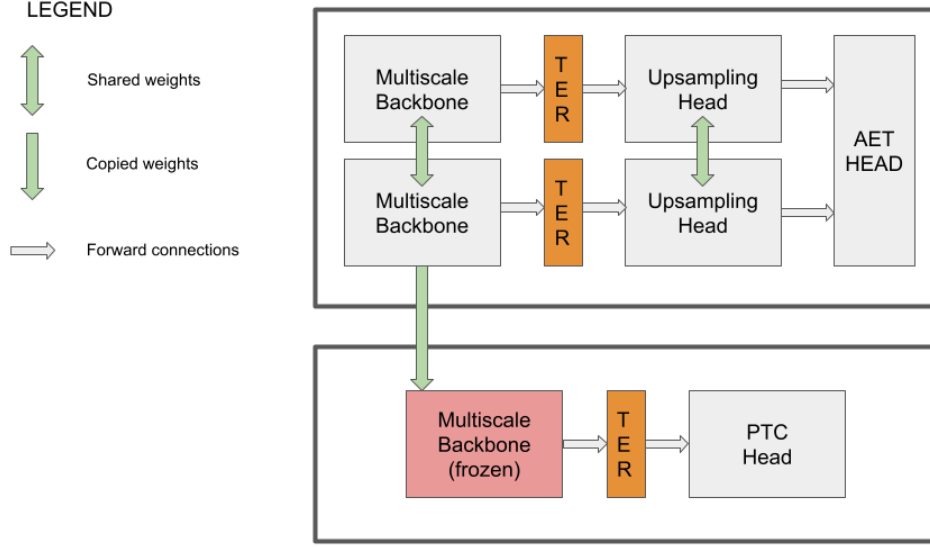


Figure 6.5: Illustration showing how key architectural elements from the AET network are incorporated in the PTC network. Red boxes indicate frozen weights.

for each batch x is sampled from three random distributions whose ranges are defined by the perceptual thresholds for a given image:

$$x \in \mathbb{R} : \begin{cases} (\log_2(0.1), x_{t-}), & \text{if } x < x_{t-} \\ (x_{t+}, \log_2(10)), & \text{if } x > x_{t+} \\ [x_{t-}, x_{t+}], & \text{otherwise} \end{cases} \quad (6.8)$$

The distribution for $c = 2$ is log-uniform, whereas the distributions for classes 0 and 1 are exponential distributions biased towards values of x lying close to the thresholds x_{t-} and x_{t+} respectively. These three values of x are then used to create three processed images and corresponding target masks Y , one for each class. For larger batch sizes, multiple images for each class are sampled. To improve generalization, image augmentation techniques are used, specifically random zooming, rotation, and cropping. These are selected as in order to keep relative pixel intensities unchanged. Horizontal and vertical flipping, as well as random scaling and cropping in the range 110-150% are each performed with 50% probability.

PTC: Objective & Optimizer Details

The optimization process adopted for the PTC is based on the approach from Section 6.3.5 with minor changes to address the size of the training dataset First, a loss function appropriate for pixel-wise classification with an imbalanced dataset is selected. In most images in the training dataset, the background class occupies more pixels than either of the suprathreshold classes. In the context of DL, such dataset imbalance is not desirable

and tends to lead to suboptimal accuracy. This imbalance can be addressed by reducing the contribution of easy classification examples (presumed to be more common in the dataset) to the overall loss. This can be done at the dataset design stage, or by modifying the training loss. Focal loss (Lin et al., 2017) is one such approach. Focal loss reduces the contribution of easy examples to the overall loss, by scaling the loss for each example based on the confidence:

$$FL(p_k) = -\alpha_t(1 - p_k)^\gamma \log(p_k) \quad (6.9)$$

where p_k is the output of the model, indicating the probability assigned to class k , α is a balancing factor and γ is the focusing parameter. In the presented experiments, these parameters are set to their default values ($\alpha = 0.25$, $\gamma = 2.0$).

The PTC model is trained with a batch size of 12, using early stopping to cease training when no improvement in validation loss is seen for 400 epochs. The model maximising the validation mean intersection-over-union measure is selected for further evaluation. In order to evaluate the relevance of the AET features, the network is also trained with progressive freezing of the AET backbone.

PTC: Evaluation Protocol

In order to evaluate the performance of the PTC, 5-fold cross-validation is used: the model is trained on 5 random folds of the training and validation data, reporting average MSE between the predicted and ground truth thresholds in the validation sets. To achieve this, a psychometrics-inspired method for finding the model’s decision boundary is developed, whereby the model’s decision boundary serves as a threshold to be compared against empirical thresholds from the perceptual experiments. Specifically, the soft F1 score between the output probability maps and the ground truth object masks is calculated for a range of transformation magnitudes. A threshold can then be defined based on the transformation magnitude corresponding to a particular F1 score. The soft F1 score is defined as:

$$F1(Y, \hat{Y}) = 2 * \frac{\epsilon + \sum_{i=0}^{H \times W} |Y_i \hat{Y}_i|}{\epsilon + \sum_{i=0}^{H \times W} |Y_i| + \sum_{i=0}^{H \times W} |\hat{Y}_i|} \quad (6.10)$$

where Y is the ground truth transformation mask, \hat{Y} is the predicted transformation mask, ϵ is a small constant. When evaluating the AET for each image from the validation dataset, a threshold is placed at the transformation magnitude corresponding to an F1 score of 0.1.

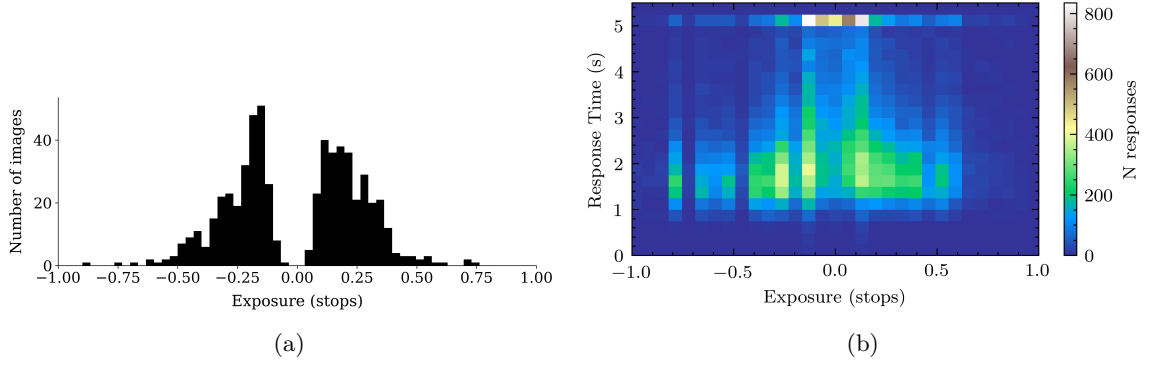


Figure 6.6: Illustration of experimental results: a) Empirical thresholds collected in the 2AFC experiment, as well as b) corresponding response time distribution, expressed as a function of exposure offset

6.4 Results

This section presents the results of the 2AFC study, as well as a performance evaluation of both the AET and PTC models. In particular, a baseline evaluation is performed on each of the models. This is followed by extended evaluation of the PTC, both against ground truth JNDs, as well as in the context of auxiliary tasks, such as composite object localisation. In addition, the feature embeddings learned by the AET are compared to baseline ImageNet-based embeddings to illustrate the transformation equivariance of the proposed model. A cross-validation of the PTC is also presented, indicating its superior performance on the perceptual threshold dataset, compared to baseline methods. Finally, the outputs of both models are illustrated, and additional qualitative evaluation is performed using an authentic image composite dataset from Xie, Xu and Chen (2012).

6.4.1 Perceptual Threshold Estimation

In the 2AFC study, a total of 41725 unique responses are obtained, with an average of 23.14 responses per observer-image combination. A total of 590 mean thresholds for 295 images are calculated after fitting psychometric functions, bootstrapping and removing images with outlier thresholds beyond 3 standard deviations (Fig. 6.6a). The means of the resulting threshold distributions are $x_{t-} = -0.2478$ and $x_{t+} = 0.2280$ for negative and positive thresholds, respectively. Observers take on average 2.65s per response. Analysis of distributions of response times and exposure transformation magnitudes sampled by the QUEST procedure (Fig. 6.6b) shows that response times (RTs) fall into two distinct groups - responses provided *within* and *outside* the 5s time limit. In the former scenario, it can be observed that shorter RTs are more commonly observed when the transformation magnitude is high, while lower-magnitude transformations tend to contribute to higher RTs.

On average, perceptual thresholds were lower for highly-textured and bright objects. Significant correlations between the mean luminance of target objects and corresponding

mean thresholds were also found. For negative offsets, the Pearson product-moment correlation coefficient was $r = .25$ ($p \leq .001$) and $r = -.39$ ($p \leq .001$) for positive offsets. A similar correlation between the standard deviation of object luminance values was found: $r = .30$ ($p \leq .001$) for negative and $r = -.45$ ($p \leq .001$) for positive offsets. No significant correlations between perceptual thresholds and target object areas were observed. However, the highest perceptual thresholds were observed in images with very small objects. This is illustrated in Figure 6.7, which shows examples of mean perceptual thresholds visualised for individual images, based on the results of the 2AFC study. Figures 6.7a-6.7d show images, where positive and negative perceptual threshold values are balanced, Figures 6.7e-6.7h illustrate images with the highest positive thresholds, while Figures 6.7i-6.7l show examples of images with the highest negative thresholds. In these examples, small dark objects, particularly those with weak textures, are often associated with high perceptual thresholds. For example, the vase in Figure 6.7f requires a 0.74 stop exposure shift before being reliably detected as modified. Also, objects in visually busy scenes, or occluded locations, tend to yield higher thresholds, like the box of chocolates in Figure 6.7h, the rear car in Figure 6.7j, or the laptop in Figure 6.7l.

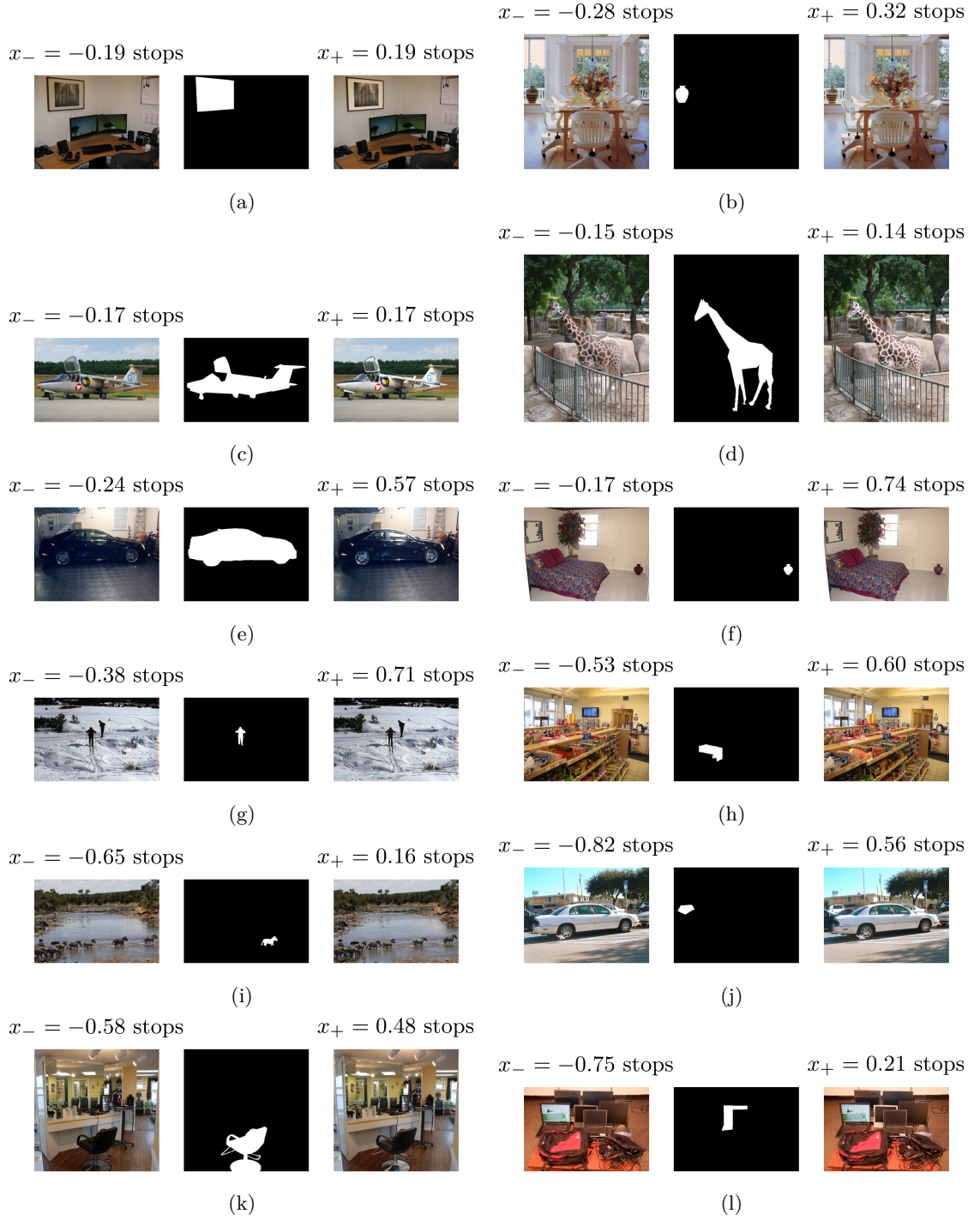


Figure 6.7: Visualisation of mean image-wise perceptual thresholds collected in the 2AFC study. Subfigures a) through d) show examples of balanced positive and negative thresholds, located near the mean thresholds of the dataset. Subfigures e) through h) illustrate examples of images with the highest positive threshold values, while subfigures i) through l) show examples of images with the highest negative threshold values.

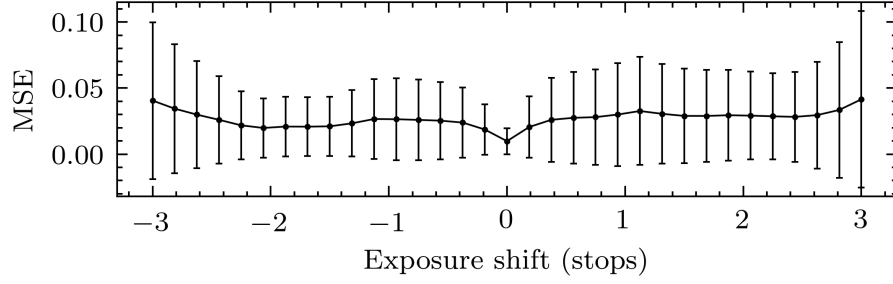


Figure 6.8: Mean MSE between ground truth and prediction for AET prediction errors for the validation dataset, across a range of exposure shifts. Error bars indicate the standard deviation of MSE for each transformation magnitude.

6.4.2 AET

The AET is evaluated by calculating MSE between predicted \hat{Y} and ground truth Y transformation masks for a range of transformation magnitudes. Figure 6.8 illustrates mean errors and their standard deviations across the 5000 validation images, for a set of 33 discrete transformation magnitudes between -3 and $+3$ exposure stops. Overall, as the magnitude of the transformation is increased, the AET’s prediction accuracy decreases. This is likely due to the clipping that can occur at high magnitudes, reducing the relative differences between neighbouring pixels and consequently removing texture information. The model achieves the lowest average errors when the transformation parameter is 0.

Figure 6.9 illustrates the output of the trained AET for multiple exposure offsets applied to the same input image. The line plot in Fig. 6.9e illustrates the normalised MSE between the ground truth and predicted transformation masks. It can be seen that the prediction error increases as a function of transformation magnitude. The heatmaps in Figures 6.9c and 6.9d show the predicted and ground truth transformation maps, respectively. Green indicates positive exposure shifts, pink indicates negative exposure shifts.

Figure 6.10 illustrates the impact of the proposed pre-training procedure on the distribution of image embeddings in feature space. Each of the subfigures illustrates the embeddings of 11 variations of an image generated by applying exposure shifts of different magnitudes (annotated for each point). Embeddings in the left figure are generated using a Resnet-50 network pretrained on ImageNet, while the central and right plot are generated using the AET’s bottleneck and output features. For each model, features for 100 images from the validation set are first extracted and PCA is then performed on those features for each model. It can be seen that both the *AET latent* and *AET output* features encode the direction and magnitude of exposure transformations applied to the input better than the Resnet-50 baseline.

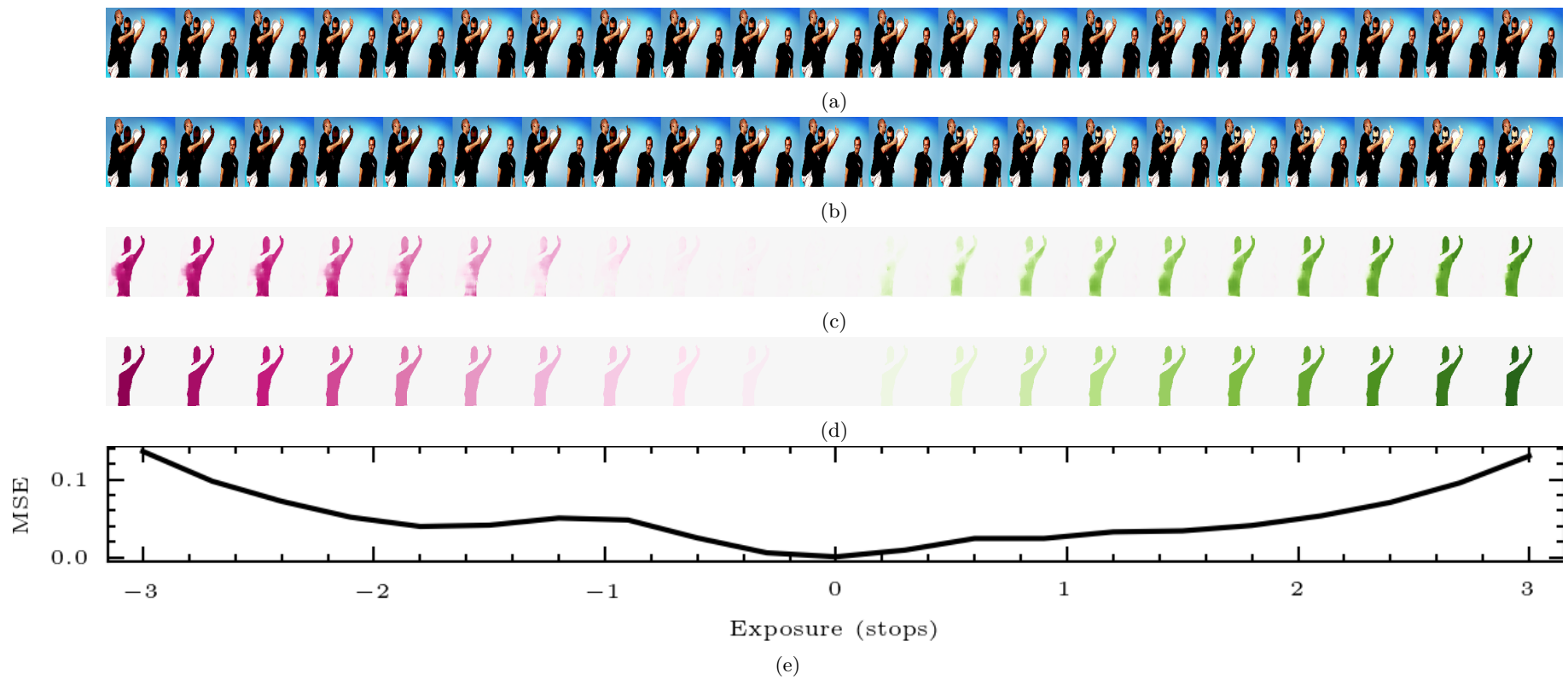


Figure 6.9: Illustration of AET output vs ground truth for a series of inputs. a) original images and b) synthetic composite images affected by local exposure transformation. c) predicted transformation maps, d) ground truth transformation maps, e) MSE between ground truth and predicted transformation maps, illustrated for a range of exposure shift values.

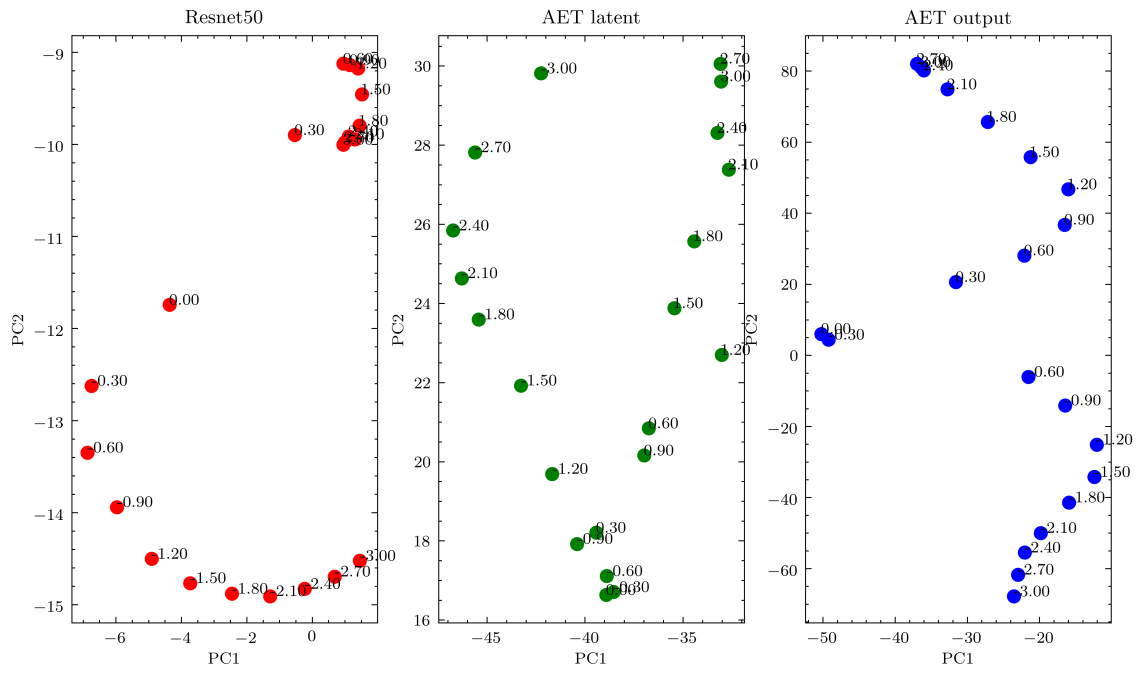
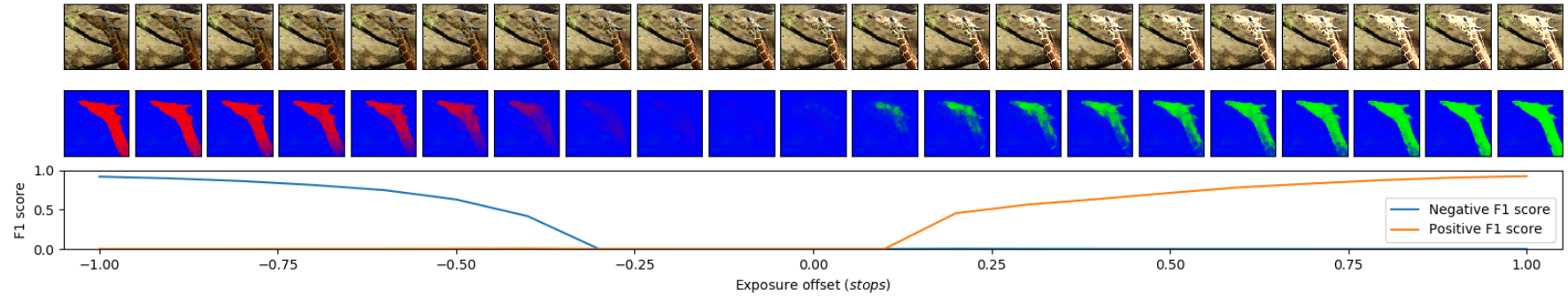
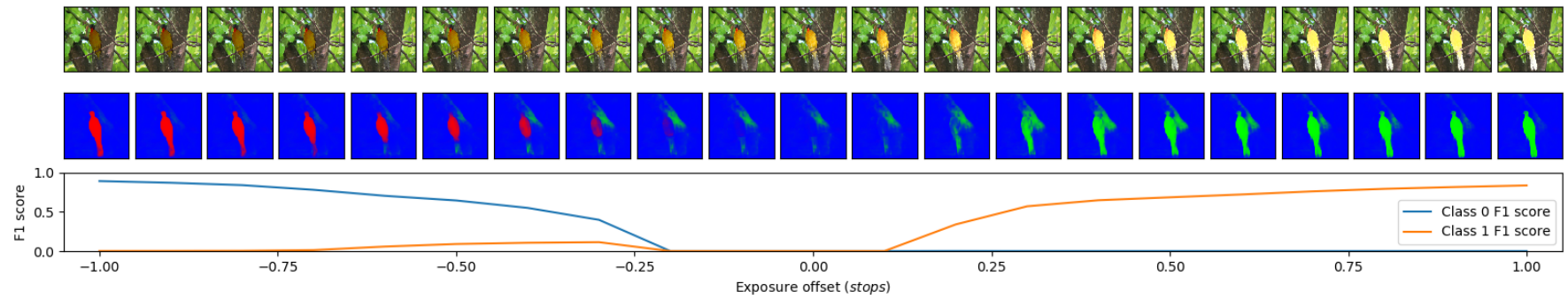


Figure 6.10: Principal component visualisation of features extracted from images affected by local transformations of different magnitudes. *Left:* Resnet50 pretrained on ImageNet - output features, no pooling; *centre:* AET - bottleneck layer features; *right:* AET - output layer features

6.4.3 Perceptual Threshold Learning

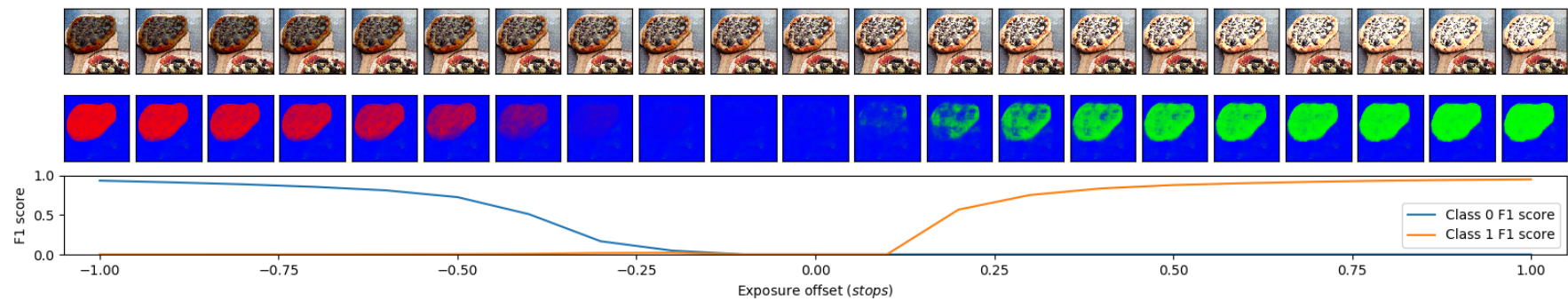


(a)

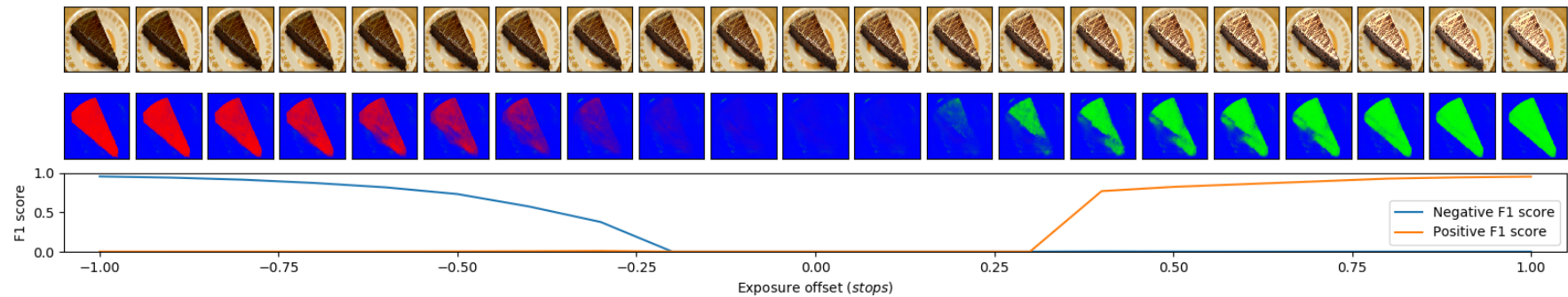


(b)

Figure 6.11: Illustration of how change in $F1$ score between *predicted* and *ground truth* (not shown here) masks is used to estimate our model's decision boundary. The top row shows input images, the middle row shows model prediction softmax probabilities with **red** for detected negative offsets (class 0), **green** for positive offsets (class 1) and **blue** for no offset. The bottom row shows class-wise $F1$ scores for classes 0 and 1. More examples in Figure 6.12.



(a)



(b)

Figure 6.12: Illustration of how change in $F1$ score between *predicted* and *ground truth* (not shown here) masks is used to estimate our model's decision boundary. The top row shows input images, the middle row shows model prediction softmax probabilities with red for detected negative offsets (class 0), green for positive offsets (class 1) and blue for no offset. The bottom row shows class-wise $F1$ scores for classes 0 and 1.

Freeze Up To Layer	MSE both	MSE x_{t-}	MSE x_{t+}
no freeze	3.9690	3.5716	4.3664
block1 pool	0.3028	0.2618	0.3442
block2 pool	0.2098	0.2188	0.2000
block3 pool	0.1895	0.1633	0.2161
block4 pool	0.2350	0.2025	0.2681
block5 pool	0.1335	0.1624	0.1046
concatenate	0.1148	0.1307	0.0978

Table 6.1: Cross-validation results: Average mean squared validation errors between ground truth thresholds and model predictions are given in exposure stops. Individual errors for positive and negative exposure offsets are shown in the rightmost two columns. Errors in each row are a result of freezing progressive parts of the pre-trained AET backbone.

As no previous work has addressed the problem of perceptual threshold approximation, it is impossible to compare the proposed model to existing approaches. Consequently, See Figure 6.11 for an illustration of the soft F1 score as a function of exposure shift. More visual examples can be found in the supplementary materials.

To evaluate the relevance of features learned by the AET, this analysis is performed for a range of fine-tuning regimes, where different parts of the model are frozen before training. The results of this experiment can be seen in Table 6.1. Overall, the results indicate the benefits of adopting both the AET and multiscale extension, particularly considering the performance increase afforded by freezing the entire encoder and only fine-tuning the decoder. The model’s performance drops significantly when the pre-training stage is omitted or when all layers of the pre-trained model are allowed to be fine-tuned.

6.4.4 Application to Real Composite Images

In order to evaluate the generalisability of the PTC to other transformation types and non-synthetic (as in: not generated by applying transformations to real images) image composites, the model is applied to the dataset proposed by Xue et al. (2012). As the composites in this dataset contain elements from different source images, there are no transformation magnitudes that could be compared against the empirical JNDs from the experiments presented here. Instead, the normalised spatial response of the model is visualised to verify whether it can localise composited elements accurately. To perform this, the pixel-wise maximum suprathreshold prediction is taken, as differentiation between positive and negative transformations is not needed for this evaluation

$$P_{i,j} = \max(\hat{Y}_{i,j,p-}, \hat{Y}_{i,j,p+}) \quad (6.11)$$

Here, $\hat{Y}_{i,j,p-}$ is the predicted probability of the negative suprathreshold class for the pixel

located at the i th row and j th column of the input image, \hat{Y}_{i,j,p_-} is the respective predicted probability of the positive suprathreshold class, and P is a matrix containing a single probability value for each pixel of the input image. This is followed a normalisation step

$$\hat{M}_{i,j} = \frac{P_{i,j}}{\max(P)} \quad (6.12)$$

resulting in the normalised matrix \hat{M} . See the right-most images of Figures 6.13 and 6.14 for visual examples of \hat{M} .

In order to quantify the performance on true (as opposed to synthetic) image composites the receiver operating characteristic (ROC) is calculated (see Fig. 6.15 for illustration). The ROC curve illustrates the change in the true positive and false positive rates of the classifier as the discrimination threshold is varied.

6.5 Discussion

The results of this study indicate that the methodology proposed in this chapter is an effective approach to both detecting effects of local transformations in natural images, as well as learning generalisable observer functions, describing subjective properties of natural images. This can be learned directly from image data and corresponding subjective responses, or models thereof. This section discusses each element of this study, as well as the obtained results, in the context of preceding chapters and related work.

6.5.1 2AFC Study & JNDs

Compared to the results from Chapter 4, where the generalised JNDs for exposure transformations were -0.54 stops $95\% CI[-0.37, -0.64]$ and 0.30 stops $95\% CI[0.13, 0.38]$ for negative and positive exposure transformations respectively, the mean JNDs obtained in this experiment were lower (-0.25 and 0.23 stops respectively). Since the QUEST method used in the latter study presented observers with many sequential repetitions of the same image, while the former only displayed each scene once per observer, such a tightening of JNDs is expected. Additionally, the dataset for this study was expanded, compared to that in Chapter 4. Finally, since experiments were conducted in an image-wise fashion, the strategy for aggregation of the 2AFC data is also different: the responses of observers for a *single* image were aggregated (as opposed to aggregating across all images), followed by the fitting of a psychometric function using the methodology from Chapter 4. The JNDs obtained in this manner are also more balanced, compared to those from Chapter 4, where JNDs for negative exposure transformations were considerably higher than those for positive transformations. These results suggest that image-wise calculation of JNDs enables implicit modelling of the contextual factors discussed in Chapters 4 and 5, such as the base object appearance, its immediate surroundings, as well as auxiliary



Figure 6.13: Input composite images from Xue et al. (2012) (left), ground truth composite masks (middle) and the normalised outputs of the PTC for suprathreshold classes (right). Despite being trained solely on exposure shifts, the model generalises well to detecting appearance differences present in real composites.

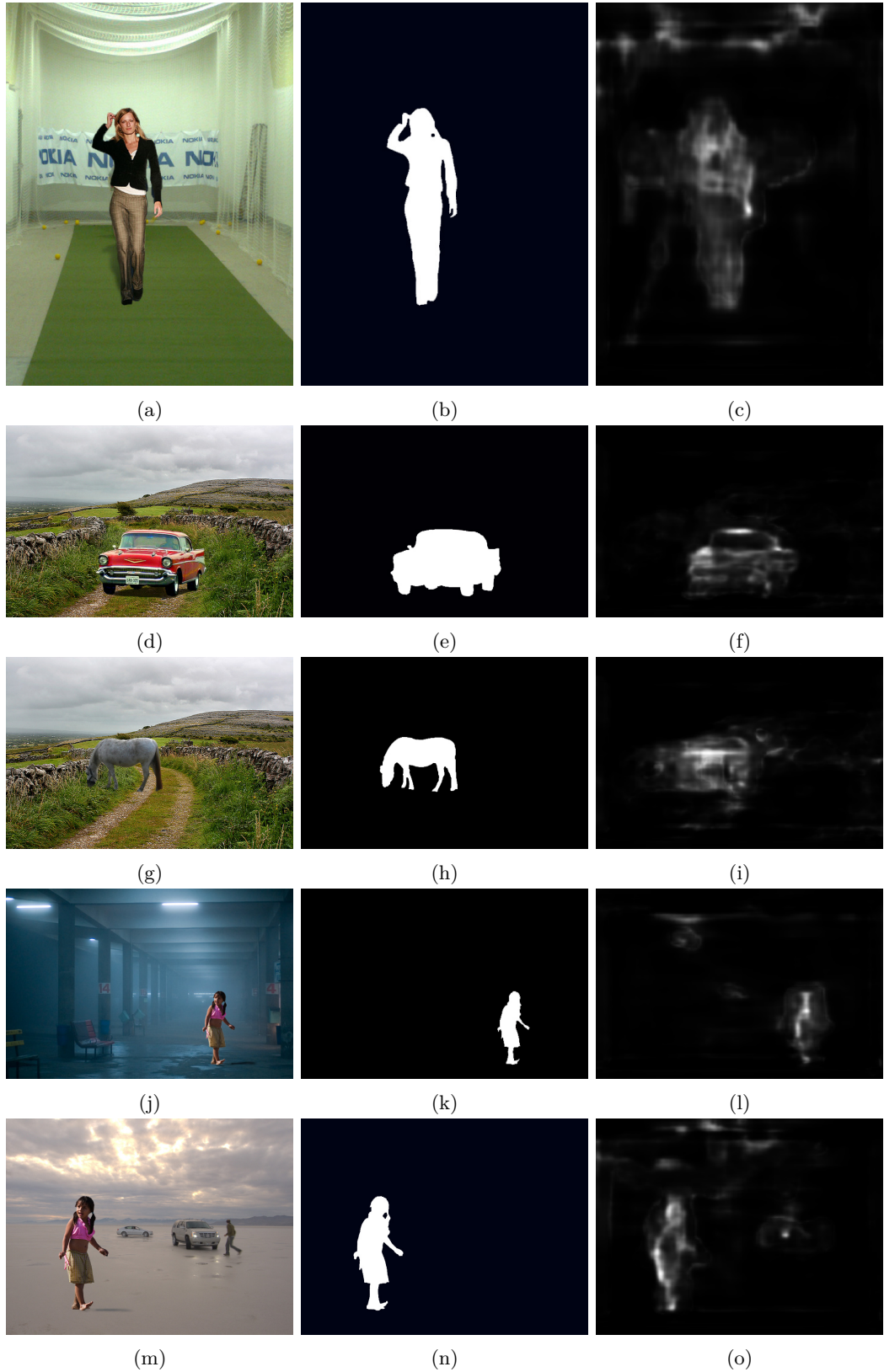


Figure 6.14: Input composite images from Xue et al. (2012) (left), ground truth composite masks (middle) and the normalised outputs of the PTC for suprathreshold classes (right). Despite being trained solely on exposure shifts, the model generalises well to detecting appearance differences present in real composites.

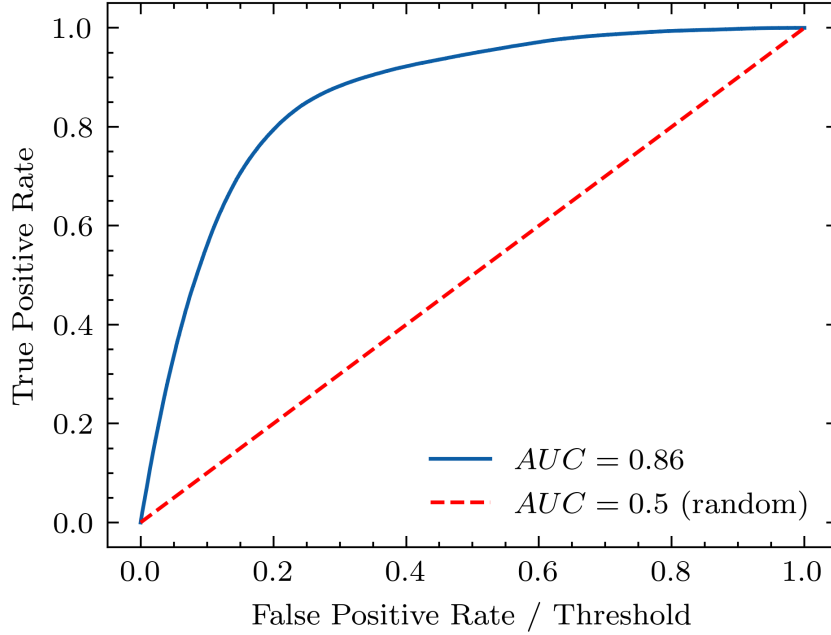


Figure 6.15: ROC curve illustrating the performance of the *PTC* when used for localisation of composited objects in the dataset from [Xue et al. \(2012\)](#). The abscissa illustrates both the false positive rate, as well as the discrimination threshold, for which both rates are calculated. The red dashed line illustrates the performance of a random classifier for comparison. The AUC label in the legend illustrates the area under the curve.

factors such as nearby illumination. This is enabled by the nature of the proposed model, which creates a perceptual mapping for every pixel of the input.

6.5.2 TER via AET

The effectiveness of the TER, learned following the AET paradigm, is best illustrated when applying it to an auxiliary task - in this case as a feature extractor in a perceptual classification task using the *PTC*. As illustrated in [Table 6.1](#), the *PTC* achieves the best performance when the weights of the AET are not allowed to be further tuned. As such, the training task is reduced to learning a decoder to map from the TER feature space to the space of the perceptual task, as defined in [Section 6.3](#). This also indicates that the TER is able to encode the approximate location and magnitude of a transformation, from local appearance statistics, without relying on a reference image.

These results also reinforce the argument of [Bengio, Courville and Vincent \(2013\)](#), who stressed the importance of data representation in machine learning problems. For example, it is well-known that classifiers and object detectors perform best when relying on feature representations invariant to appearance differences ([Schmidt and Roth, 2012](#)). On the other hand, the problem of visual realism, or other subjective properties based on distortion visibility, requires these distortions to be well represented in feature space, if they are to be successfully mapped to a subjective opinion score, or saliency map. In the presented methodology, this transformation equivariance is encoded through self-supervised training

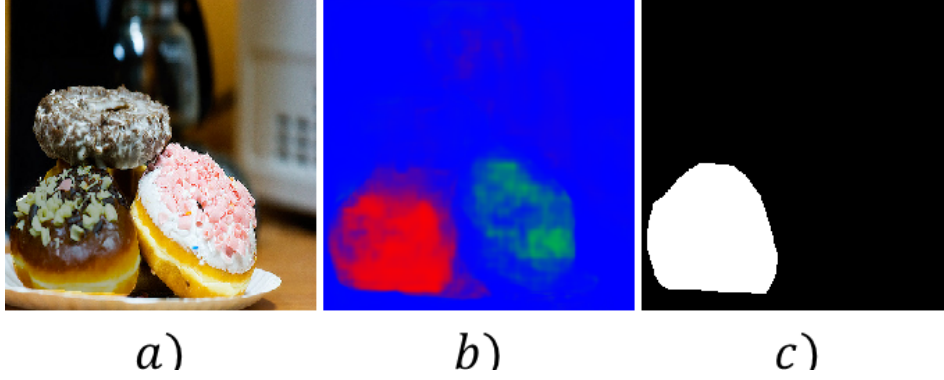


Figure 6.16: Example of *a)* Over-exposure resulting from flash or spot lighting in the original image *b)* both the original over-exposure (green) and manually applied underexposure (red) are detected by our model *c)* mask showing area where negative exposure shift is manually applied

and a relevant data augmentation strategy. However, this is not to say that more optimal solutions do not exist.

While more efficient implementations of the proposed methodology and various optimisations are certainly possible, the proposed framework is general-purpose and can be easily applied to various local image transformations, which can be annotated using a pixel-wise parameter map, as illustrated in Section 6.3.

6.5.3 Approximation of Empirical JNDs

In terms of alignment with experimental data, the PTC obtained a low average *MSE* of 0.11 stops when comparing predicted to ground truth JNDs. Despite the fine-tuning of the PTC being performed using a relatively small training set, the model performs well on unseen data. For example, in Figure 6.16 the model detects both the transformation applied to it, as well as the pre-existing over-exposure of the object on the right. A similar scenario was illustrated in Figure 6.11b, however, in this scenario the change in the appearance of the foreground object triggers a false positive. Interestingly, this only occurs for a select few transformation magnitudes, suggesting the potential of reciprocal interactions between local image regions, whereby a change in one impacts the model’s prediction of a directly neighbouring region. This also illustrates a wider problem, whereby high-contrast regions in natural images may trigger false positives. Further analysis is necessary to understand the causes behind such edge cases.

6.5.4 Towards a General Representation of Naturalness

In addition to the effectiveness of the learned representation on the task of local transformation detection in synthetic image composites, there is evidence that these features, conditioned on exposure transformations, are at least partly generalisable to other image composite artefacts. This is illustrated when applying the PTC to authentic image composites, featuring objects from other images, rather than just local transformations. As seen in Figures 6.13 and 6.14, the PTC is capable of localising these objects consistently

(see Fig.6.15) on unseen, out-of-domain data, without any additional training. This also signifies that exposure shifts may perceptually correlate with the types of distortions present in authentic composites. These likely encode more complex appearance differences, which are a result of a combination of multiple transformations, some of which may be nonlinear. Consider a combination of some effects discussed in Chapter 2, such as differences in scene illumination or in-camera processing. While their combined effect may be a nonlinear function of the input, perhaps the statistical differences between object and scene induced by a local exposure transformation are sufficiently similar to those present in an authentic composite.

Consequently, assessment of visual realism in image composites may not necessarily require a TER trained on an exhaustive range of specific transformations. Instead, this problem can be abstracted to the detection of any and all image perturbations, followed by perceptual tuning, using a small subset of subjectively labelled data. As such, the TER would have to learn to encode the local statistical ‘naturalness’ of image regions, as well as, more importantly, the local deviation away from this naturalness. To achieve this, one could rely on the same set of self-supervised training techniques and generalise the transformation parameter maps to logarithmic difference maps.

6.5.5 Wider Applications

Identification of unrealistic composited objects can be applied directly to image compositing and harmonisation problems. State-of-the-art approaches to composite realism improvement, such as Sunkavalli et al. (2010) or Tsai et al. (2017), often require explicit identification of pixels belonging to the composited object by specifying binary object masks, thus significantly limiting application scenarios to ones where such object masks are available. As existing harmonisation models also rely on CNN-based architectures and gradient-based optimisation techniques, the model proposed in this chapter could be incorporated into such harmonisation pipelines and address this limitation with little modification. This, in turn, would allow for no-reference harmonisation of image composites, using the output of the PTC as a replacement for a pre-existing ground truth object mask.

6.6 Conclusions

This chapter has presented a novel methodology for the detection of local suprathreshold image transformations based on approximation of the image-wise function performed by an observer. This is achieved by first learning a transformation equivariant representation conditioned on local exposure transformations, followed by fine-tuning of a fully convolutional image classifier on top of the TER, and conditioning its class decision boundaries using a data generation scheme based on empirical perceptual thresholds corresponding to JNDs. The resulting model is capable of approximating

human JNDs with an average error of 0.11 stops. Furthermore, when applied to authentic image composites, the proposed model is able to reliably indicate the location of the composited object, suggesting that the features learned from training on local exposure transformations applied to natural images can be generalisable to composite images in the wild. While the proposed method is illustrated only for local exposure transformations, it can be applied to many local distortions or transformations, provided these can be synthesised and represented by a transformation map. A further interesting direction is the extension of this approach to high dynamic range images, as this study only evaluates 8-bit images, and thus the presented results are constrained to this bit depth.

In order to illustrate a practical application scenario, and building on the results of applying the proposed models to authentic image composites, Chapter 7 will investigate how the proposed models can be applied to the perceptually-related task of image harmonisation.

Chapter 7

Applying Perceptual Models in Image Harmonisation

This work was published in:

Dolhasz., A., Harvey., C. and Williams., I., 2020. Towards unsupervised image harmonisation. *Proceedings of the 15th international joint conference on computer vision, imaging and computer graphics theory and applications - volume 5: Visapp.*, INSTICC, SciTePress, pp.574–581. Available from: <http://doi.org/10.5220/0009354705740581>

extended version is pending publication in:

SPRINGER CCIS

7.1 Introduction

Chapter 6 illustrated that a deep convolutional neural network can be used to model human perceptual sensitivity in the context of composite artifact detection. This was achieved by learning a mapping from composite images to corresponding masks indicating the existence of local suprathreshold transformations, based on image-wise JNDs, measured using the methodology from Chapter 4 and informed by empirical evidence about spatial allocation of attention from Chapter 5. A methodology for data collection, self-supervised pre-training, and model fine-tuning based on perceptual data were proposed. Upon evaluation, the resulting model was shown to approximate image-wise perceptual thresholds with low error, as well as localising composited objects in authentic image composites, suggesting the features learned by the model may be transferable to auxiliary tasks. Consequently, this chapter combines and applies the findings thus far to address the core problem of this thesis, namely image composite harmonisation: perceptually-informed improvement of composite realism.

This is accomplished through joint and explicit modelling of both *detection* and *harmonisation* of composite artefacts in an end-to-end manner. The detection task

is performed using the model proposed in Chapter 6, while the harmonisation task is accomplished using a state-of-the-art harmonisation model. Such an approach allows for the harmonisation process to be informed directly by the perceptually-calibrated detection model, instead of relying on binary object masks required at input by existing harmonisation techniques. In combination with previous chapters, the work discussed here illustrates the final step in a general framework for perceptual calibration of image compositing tasks. This framework consists of empirical modelling of observer sensitivity to local image transformations (Ch. 4), the corresponding spatial attention allocation (Ch. 5), the generalisation of such models using transfer learning techniques (Ch. 6) and finally their integration and joint training with state-of-the-art harmonisation techniques, presented in this chapter. This novel approach is first evaluated through a baseline combination of pre-trained models and then extended into a single, end-to-end model, capable of outperforming current state-of-art techniques across two datasets and showing improvement in the compositing task, using two different end-to-end model architectures.

7.2 Related Work

7.2.1 Image Compositing, Harmonisation and Deep Learning

As illustrated in previous chapters, a naïve combination of elements from different source images is very likely to produce an unrealistic compositing result, due to the various appearance-based differences present between natural scenes, which become noticeable and negatively impact visual realism when these elements are combined. To address this, such appearance-based differences between individual elements of a composite should be minimised, in order to produce a plausible end result (Wright, 2013b). As such, harmonisation is perhaps the most important stage of the compositing process, when a high degree of visual realism is desired.

Similarly to the problem of image in-painting (Bertalmio et al., 2000) or the extraction of 3-D information from a 2-D retinal image (discussed briefly in Chapter 2), compositing and harmonisation are both ill-posed problems (Guillemot and Le Meur, 2013). In contrast to problems where the solution is unique, for a given region of an image composite requiring correction, many different arrangements of pixels could be deemed plausible. Depending on the content and context of a composite, some scene properties, and thus required object corrections, may be inferred from the information contained within the image or its metadata, such as the characteristics of the illuminant (Shi, Loy and Tang, 2016), colour palette, contrast range or the camera response function. Other properties, such as an object’s albedo, texture or shape are often unique to the object and cannot be derived directly from contextual information in the scene. While methods for approximation of these properties do exist (Gardner et al., 2017), they are difficult to integrate into end-to-end systems and can be challenging to parametrise. Recently, advances in DL (see Chapter 3 for a review) have motivated a number of approaches which exploit

the huge amount of natural imagery available in public datasets in order to learn a mapping between corrupted composite images and their corrected counterparts, or natural images. Several methods relying on variants of convolutional autoencoders (AEs) have been successfully used to directly approximate the harmonisation function, in a supervised learning setting. Notably, Tsai et al. (2017) use a convolutional AE in a multi-task setting to both segment and harmonise an input image, provided the target object mask. Chen and Kae (2019) use a generative adversarial network (GAN) to perform both colour and geometric transformations, pre-training their model on synthetically-generated data. Conditional GANs have also been applied in this context (Azadi et al., 2018), by learning to model joint distributions of different object classes and their relationships in image space. This allows for semantically similar regions to undergo similar transformations. Another technique, proposed by Cong et al. (2020), combines attention mechanisms and GAN-based architectures with explicit object-scene knowledge implemented through masked and partial convolutions and provide a dedicated benchmark image harmonisation dataset, dubbed iHarmony.

However, as mentioned in Section 6.5, a common requirement of these existing techniques is the provision of binary object/scene segmentation masks at input, both during training and inference. These masks serve as an additional feature, or a form of manually encoded attention, identifying the image pixels that require harmonisation. As such, these methods are only applicable to scenarios where new composites are generated, and these masks are available. In cases where these ground truth masks are not available, these techniques can not be applied without human intervention, precluding their application to scenarios such as harmonisation of legacy composites. Moreover, existing methods do not explicitly leverage perceptual characteristics of humans - the conventional target audience for image composites. Human sensitivity to different local image disparities between object and scene has been shown to correlate with subjective realism ratings, as discussed in Chapter 4. Lastly, binary object masks used in these techniques provide only limited information about the nature of the required corrections, indicating only the area where corrections are needed. Due to the varying perceptual impact of distortions applied to regions of different appearances, as illustrated in Chapters 4 and 5, *explicit* modelling of both detection and correction of image composite artefacts may afford improvements in terms of quality, as well as allow for wider application of such harmonisation algorithms.

7.2.2 Multi-task Learning, Feature Sharing & Attention

Due to the abundance of natural image data and the ill-posed nature of the compositing problem, DL approaches are well-suited for this task. However, supervised DL methods require large amounts of annotated data in order to learn and generalise to unseen data. This requirement grows along with the complexity of a problem and the desired accuracy. In order to tackle this issue, numerous architectural considerations have been proposed, many of which focus on learning good feature representations, which generalise well

between tasks.

Multi-task learning approaches rely on performing multiple related tasks in order to learn better feature representations. In recent years many tasks in image understanding have achieved state-of-the-art performance by incorporating multitask learning (Evgeniou and Pontil, 2004), for example in predicting depth and normals from a single RGB image (Eigen and Fergus, 2015), detection of face landmarks (Zhang et al., 2014) or simultaneous image quality and distortion estimation (Kang et al., 2015). This is afforded by the implicit regularisation that training a single model for multiple related tasks imposes (Caruana, 1997), and the resulting improved generalisation. Feature sharing approaches combine deep features from related domains or tasks in order to create richer feature representations for a given task. This is similar to the multitask paradigm, however instead of sharing a common intermediate feature representation, features from one or multiple layers of two or more networks are explicitly combined. The Deep Image Harmonisation Tsai et al. (2017) (DIH) model Tsai et al. (2017) adopts both these paradigms, by combining the tasks of image segmentation and harmonisation and sharing deep features of both task branches. Finally, attention mechanisms Cun and Pun (2020) can also be used to learn the relative importance of latent features for different combinations of task and input sample.

7.2.3 Attention instead of masks

State-of-the-art image harmonisation methods focus largely on improving composites in scenarios where the identity of pixels belonging to the object and scene are known a priori. For example, the DIH approach (Tsai et al., 2017) uses an AE-based architecture to map corrupted composites to corrected ones, incorporating a two-task paradigm, which attempts to both correct the composite, as well as segmenting the scene. However, this approach does not explicitly condition the network to learn anything more about the corruption, such as its magnitude, type or location. Instead, object location information is explicitly provided at input, using a binary mask. A similar approach (Chen and Kae, 2019) inputs the object mask at training time, while also introducing mask segmentation and refinement within a GAN architecture, in addition to learning of geometric transformations of the object. The segmentation network, as part of the adversarial training process, discriminates towards ground truth binary masks as an output - omitting any perceptual factor in the discrimination task. This achieves improved results compared to the AE, however at the cost of a more complex architecture and adversarial training. Due to the many dimensions along which combinations of object and scene may vary, compositing systems should be equipped to encode such differences before attempting to correct them. Kang et al. (2015) show that a multitask approach is an efficient way to ensure that distortions are appropriately encoded by the model. Other approaches to this problem include self-supervised pre-training to enforce equivariance of the latent representation to certain input transformations (Zhang et al., 2019), which has been used to train perceptually-aligned local transformation classifiers in Chapter 6.

Additionally, as discussed in Chapter 3, removing the requirement for an input mask from image composite harmonisation algorithms would allow for application at scale to legacy content. In scenarios where raw image composite elements are not readily available, this would remove the need for manual mask generation. This in turn would allow application to classic films, photographs and photo-montages and any other finished image composites, for which source materials may no longer be available. This, in turn, would allow for reductions in the cost and labour associated with restoring legacy content. Even in scenarios where masks are available, these could be perceptually re-weighted based on the approaches proposed in this thesis, before being used in the harmonisation process. In other words, harmonisation algorithms should be able to detect and estimate the magnitude of perceptually-relevant distortions or corruptions, even if a binary mask is supplied to locate them.

7.3 Methodology

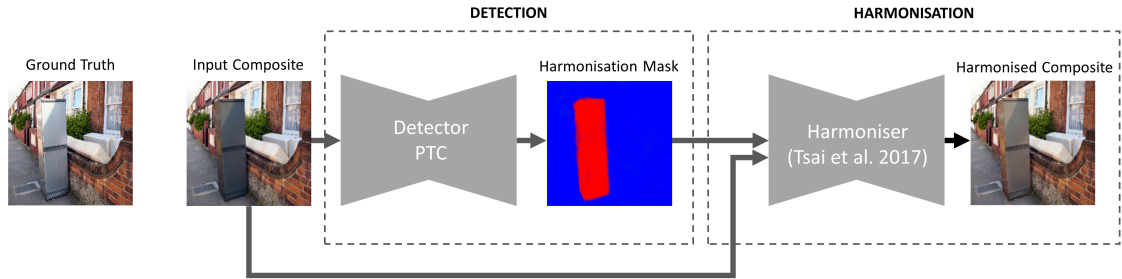


Figure 7.1: System overview: illustration of the detector and harmoniser combined into a two-stage composite harmonisation system. A synthetic composite image is first supplied to the detector, which outputs a 2-channel mask indicating detected **negative** and positive (not pictured here) exposure shifts. This mask is converted to a single-channel representation by taking a maximum over predicted pixel-wise probabilities and fed to the harmonisation network, which then produces a harmonised composite, which we compare against the ground truth.

7.3.1 Overview

The overarching goal of this thesis (stated in Section 1.3) is “*to produce perceptually-driven systems to facilitate computational subjective quality analysis of image composites and guide the subsequent manual or automated improvement of their quality.*”. Accordingly, this chapter evaluates the PTC model developed in Chapter 6 in the context of this overarching aim: perceptually-informed improvement of composite realism. Specifically, the PTC is combined with a state-of-the-art harmonisation model: Deep Image Harmonisation (DIH) proposed by Tsai et al. (2017). The authors’ original pre-trained implementation is used (see <https://github.com/wasidennis/DeepHarmonization>). The architecture of the model can be seen in Figure 3.4. In this context, the PTC is provided with an input composite image and performs the *detection* task, outputting a mask fed to the DIH model along with the input composite image, as illustrated in Figure 7.1. The modular nature of the proposed architecture allows for the component models

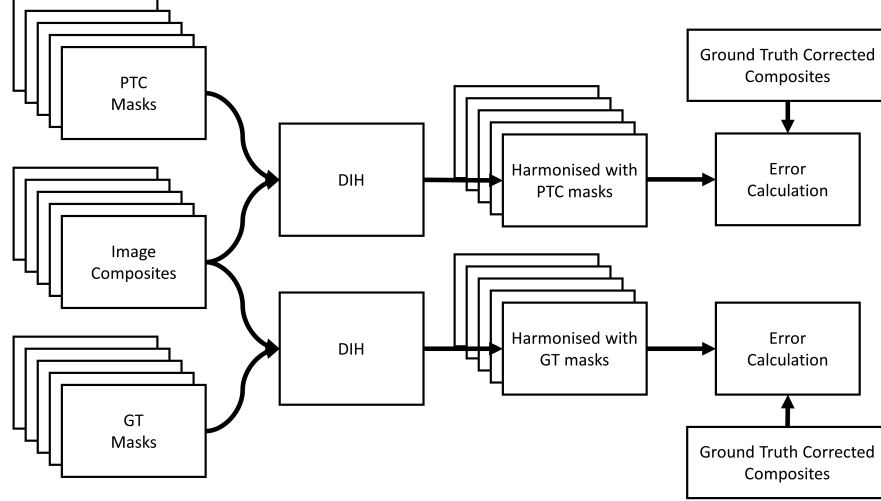


Figure 7.2: Illustration of the preliminary two-stage evaluation of the standalone models. Given a dataset of synthetic composites with corresponding ground truth (GT Masks) masks, a set of predicted masks (PTC Masks) is first generated using the PTC model. Two sets of harmonised images are then generated using the DIH model. The first set uses ground truth masks, while the second set uses masks predicted by the PTC to harmonise the input composites. Finally, errors are calculated between each of the sets of harmonised composites and the corresponding ground truth.

(the PTC and DIH) to be first evaluated separately, before being combined into an end-to-end model and trained jointly.

7.3.2 Approach

Based on the performance of the PTC on localising inserted objects in authentic composite images (discussed in Section 6.4.4), the proposed combination of the PTC and DIH can first be evaluated in a two-stage manner. This approach allows for comparison of composite harmonisation results obtained using the DIH model with ground truth binary masks against those obtained using the harmonisation masks generated by the PTC. This allows to assess whether the end-to-end process is viable using the standalone, pre-trained PTC and DIH models, without the need for additional joint training. This process is illustrated in Figure 7.2.

The hypothesis that the performance of this baseline detection and harmonisation model is comparable to a harmonisation model using manually created object masks is evaluated. Confirmation of this hypothesis would support the development and training of an end-to-end model. This research methodology is summarised in Figure 7.3.

7.3.3 Test Dataset

The test dataset is generated following the approach of Tsai et al. (2017), which is illustrated in Figure 7.4. Specifically, pairs of images (see Figs. 7.4a and 7.4c) containing objects belonging to the same semantic category are sampled from the MSCOCO dataset

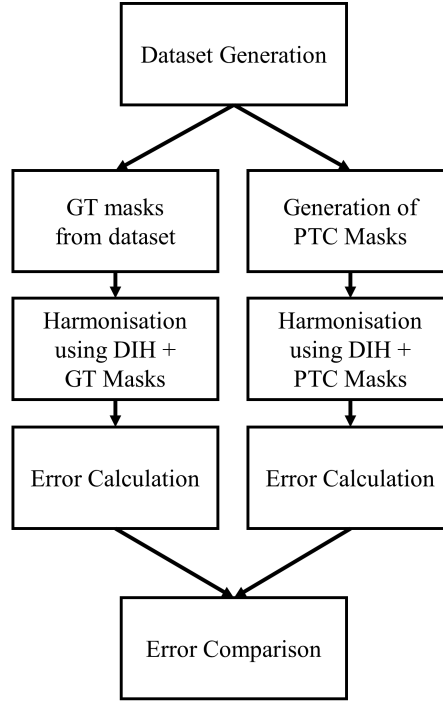


Figure 7.3: Illustration of research methodology adopted in the two-stage model evaluation.

(Lin et al., 2014), along with corresponding object masks (see Figs. 7.4b and 7.4d). Using these object masks, statistical colour transfer based on histogram matching (Reinhard et al., 2001) is performed between the pixels of objects belonging to the same category to produce a synthetic composite image (see Fig. 7.4e). Examples of these images are also shown in Figure 7.1.

Colour transfer is performed between object regions of the same semantic category. As the PTC is trained on local exposure transformations, the colour transfer is only performed on the luminance channel of the image represented in Lab colourspace.

A total of 68128 composites, corresponding ground truth images and ground truth object

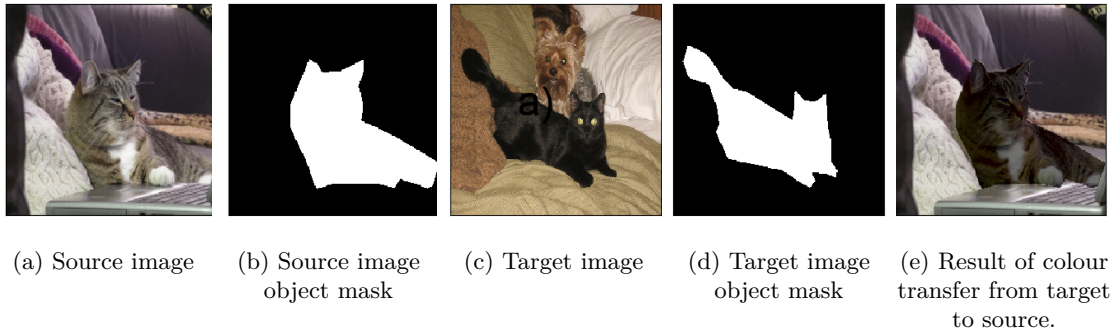


Figure 7.4: Dataset generation process adapted from Tsai et al. (2017): a) source image sampled from MSCOCO, b) corresponding object mask, c) target image, d) target image object mask, e) result of luminance transfer Reinhard et al. (2001) of source - c), to target - e.

masks are generated. For the sake of brevity, this dataset is referred to as *COCO-Exp* throughout the remainder of this chapter.

7.3.4 Evaluation Procedure

First, the input composites from the COCO-Exp dataset, along with corresponding ground truth masks are fed to the DIH model, which outputs the first set of harmonised images. Second, a set of masks is generated by feeding the input composites to the PTC. As discussed in Chapter 6 the PTC is a pixel-wise 3-class classifier and outputs maps probabilities of either the lack of, or the presence of negative or positive suprathreshold local exposure transformations for each pixel of the input image. In order to convert these into masks compatible with the DIH, the maximum of class probabilities is taken for each pixel: $M_{(i,j)} = \max(PTC_{out(i,j)})$. Here PTC_{out} is the raw output of the PTC, given an input composite, i and j are row and column indices, respectively, and M is the resulting single-channel mask compatible with the DIH. Next, the input composites are harmonised again, but this time using M instead of the ground truth masks. Finally, similarity metrics between ground truth images and composites harmonised by each of the two approaches are calculated.

7.3.5 Similarity Metrics

To evaluate each of the two approaches, similarity metrics are calculated between ground truth images and composites harmonised using the DIH model with either ground truth masks or masks predicted by the PTC, as illustrated in Figure 7.2. The similarity metrics used in this study are adopted from Tsai et al. (2017), namely Mean Squared Error (MSE):

$$MSE = \frac{1}{N} \sum_{i=0}^n (Y_i - \hat{Y}_i)^2 \quad (7.1)$$

where Y is the ground truth image and \hat{Y} is the harmonised image; and PSNR (Eq. 2.3)

here R is the maximum possible pixel intensity - 255 for an 8 bit image. In addition, the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018c) is used. This measures visual similarity based on human perceptual characteristics, providing an alternative to MSE or PSNR, which are not calibrated in line with human perception.

7.4 Results

The results of the two-stage evaluation can be seen in Figure 7.5, which shows distributions of each of the similarity metrics calculated between ground truth images and composites harmonised using the DIH model and either the ground truth masks or masks predicted by the PTC, respectively. The means of these similarity metric distributions can be seen in Table 7.1. This table shows that masks predicted by the PTC yield higher average

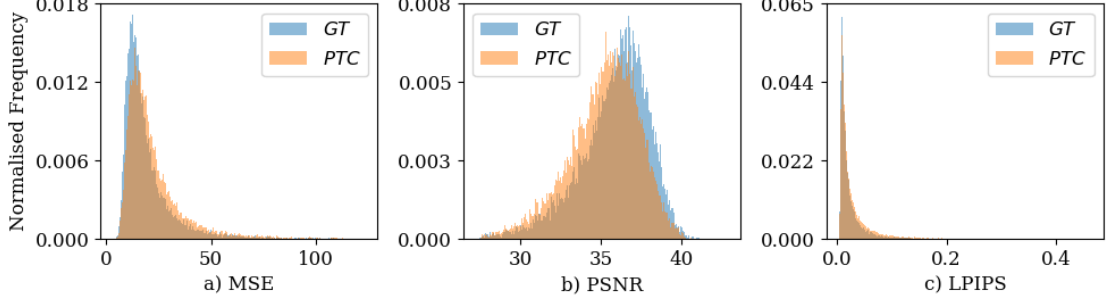


Figure 7.5: Similarity metric distributions for harmonisation using GT masks (composites corrected with synthetic ground truth masks) and PTC masks (corrected with masks predicted by the detector) (a) MSE, (b) PSNR and (c) LPIPS. Larger values indicate poorer performance for MSE and LPIPS, better for PSNR.

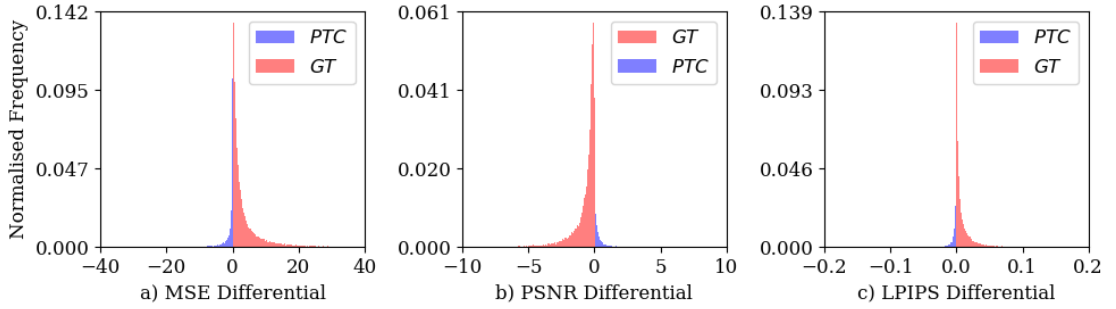


Figure 7.6: The image-wise error differentials for $C_p - C_s$, for each of the three metrics: (a) MSE, (b) PSNR and (c) LPIPS. Note, negative values for MSE and LPIPS indicate images for which C_p (composites corrected with masks predicted by the detector) achieves lower error than C_s (composites corrected with synthetic ground truth masks). For PSNR, the obverse is true.

harmonisation errors across all three metrics compared to the ground truth masks, however the magnitude of these differences is small for each of the metrics (3.1 for MSE , 0.63 for $PSNR$, and 0.0065 for $LPIPS$). Figure 7.6 shows distributions of image-wise error differentials for both techniques.

7.5 Discussion

The results obtained in the two-stage analysis indicate that using detected, instead of ground truth object masks when harmonising composites using the pre-trained DIH model

Metric	DIH + GT masks	DIH + PTC masks
MSE	19.55	22.65
PSNR	35.81	35.18
LPIPS	0.0227	0.0292

Table 7.1: Means of similarity metrics for both techniques evaluated against ground truth: DIH, and the PTC+DIH. Lower is better for LPIPS and MSE, higher is better for PSNR.

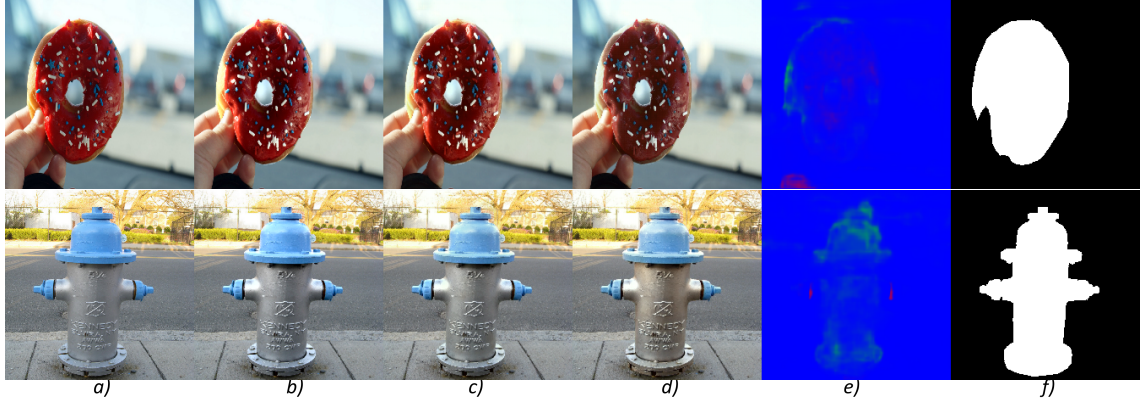


Figure 7.7: Examples of the DIH with ground truth masks over-compensating, and applying colour shifts to compensate a luminance transform, resulting in suboptimal output. From left: *a)* ground truth, *b)* input composite, *c)* output of DIH with masks predicted by the PTC, *d)* output of DIH with ground truth masks, *e)* masks predicted by PTC, *f)* ground truth masks.

results in only a slight increase in average harmonisation errors across the COCO-Exp test dataset. This suggests that additional training of the end-to-end combination of the PTC and DIH models may improve this performance, by allowing the DIH training to be influenced by the perceptually-informed PTC masks.

Importantly, as Figure 7.6 shows, harmonisation errors for a subset of the dataset were lower for the masks predicted by the PTC, indicating that in some scenarios the perceptually-informed masks can outperform ground truth masks. An example of this can be seen in Figure 7.7, which illustrates examples of failure cases, where Figures 7.7c, showing harmonisation results using predicted masks, and 7.7d showing harmonisation results using ground truth masks, illustrate a case of colour bias induced by the DIH with ground truth masks, both in the donut image as well as the hydrant image.

Further investigation indicates particular scenarios where this occurs. In some cases, the harmonisation algorithm applies an inappropriate correction, rendering a higher error for the composite harmonised using the DIH with ground truth masks, compared to the unharmonised input. Then, if the predicted masks do not approximate ground truth masks well, are blank (no detection), or their average intensity is lower than that of a corresponding ground truth mask, the additional error induced by the harmonisation algorithm is minimised, rendering lower errors for harmonisation using masks predicted by the PTC. This can be seen in both images in 7.7d. This indicates the benefit of a perceptually motivated approach to mask prediction, allowing a degree of influence over the weight of the transformation applied by the harmoniser. The DIH model tends to apply colour transformations regardless of whether they are required, based on the ground truth mask. In some cases, the perceptually-based masks produced by the PTC mitigate this problem. Images showing examples of comparable performance of the two methods can be found in Figure 7.8. Subfigures *c* and *d* show the results of harmonisation using predicted and ground truth masks respectively, and subfigures *e* and *f* show the predicted

and ground truth masks, respectively.

Due to the nature of the PTC operating solely on luminance transforms, a further benefit to the multitask learning paradigm is the generalisability to arbitrary pixel level transforms, for example colour shifts. The binary masks accepted by harmoniser networks currently do not separate across these transforms, they treat them all homogeneously. A perceptually motivated approach to the predicted mask can encode, on a feature-by-feature basis, the perceptual likelihood of harmonisation required. This is not to say, necessarily, that deep harmonisation networks cannot learn this behaviour, but provision of further support to encode this non-linearity at the input to the network, and/or by explicit optimisation at the output, would likely benefit performance and improve generalisation Caruana (1997). This is conceptually similar to curriculum learning improving convergence in reinforcement learning problems Bengio et al. (2009), or unsupervised pre-training techniques improving convergence in general.

The following section describes the development and training of an end-to-end combination of the PTC and DIH models.

7.6 End-to-End Model: Methodology

Section 7.4 illustrated that perceptually-based detection of local image transformations can be leveraged to generate masks capable of guiding image harmonisation and achieving comparable results to ground truth masks, when evaluated on an image harmonisation task using a state-of-the-art model. While harmonisation using ground truth masks achieved lower errors on average, the masks predicted by the PTC were able to mitigate the need for provision of input masks, at the cost of slightly higher harmonisation errors.

Given that this was achieved with no additional training indicates that an end-to-end model combining both these tasks could be used to perform *no reference* harmonisation, removing the need for provision of object masks for both training and inference, as opposed to current state-of-the-art approaches. Additional joint training could also allow for overall performance improvements and enable different combinations of the source models to be evaluated. Thus, to allow for fair evaluation, the DIH model is re-implemented in Tensorflow and trained anew on the iHarmony dataset (Cong et al., 2020), before being combined in an end-to-end manner with the PTC and fine-tuned. This way both the DIH and the end-to-end combination, dubbed *PTC+DIH*

7.6.1 Model Architectures

The end-to-end model is designed by combining the DIH and PTC models. First, the DIH model is implemented in Tensorflow, according to the authors' specification, and random initialisation is performed on all layers. One outer layer of the encoder and decoder in the DIH model is removed, following Cong et al. (2020), in order to accommodate for the

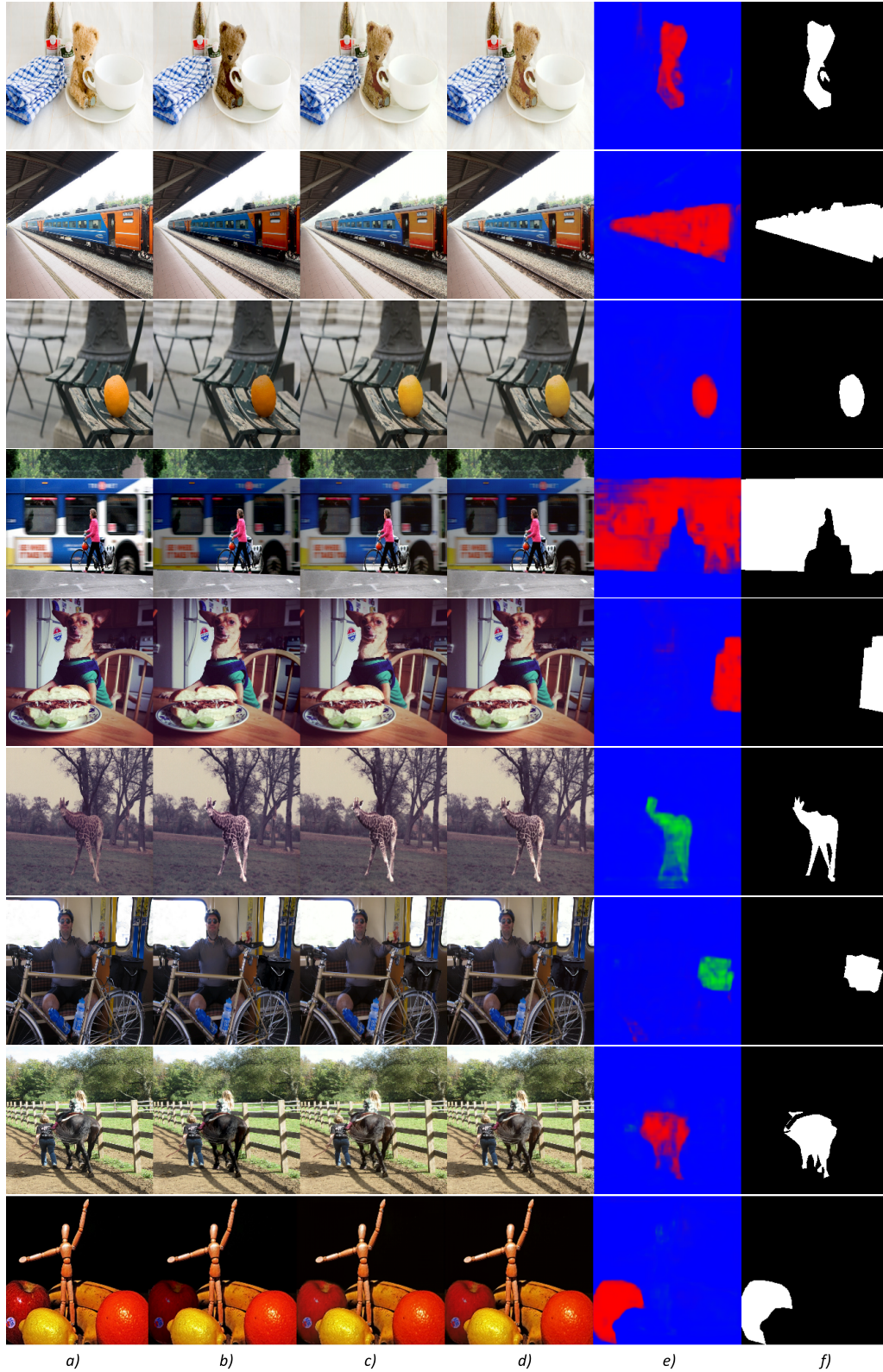


Figure 7.8: Comparison of harmonisation outputs from the evaluation. From left to right: a) ground truth, b) input composite, c) harmonised with DIH + GT masks, d) harmonised with DIH + masks predicted by PTC, e) Predicted masks, f) ground truth masks. Masks in colour indicate the raw output of the PTC, where the direction of detected luminance shifts is indicated - red for negative and green for positive shifts.

lower resolution of the PTC. All training is then performed using a resolution of 256×256 . Two approaches to combining the DIH and PTC are evaluated.

PTC+DIH

The first approach, *PTC+DIH* combines the models sequentially, whereby the PTC generates a mask from the input image, which is then concatenated with the input and fed to the DIH model, as illustrated in Figure 7.1. The original 3-class softmax output of the PTC is removed, and replaced with a single-channel sigmoid output, to match the mask input dimensions of the DIH model. Up- and downsampling operations are also added in order to adapt the input image to the 224×224 input resolution of the PTC, and its output to the 256×256 input resolution of the DIH.

PTC+att+DIH

The second approach, *PTC+att+DIH*, inspired by self-attention mechanisms (Vaswani et al., 2017), relies on combining the latent features of both models through an attention-like dot product:

$$a_{joint} = fc_3\left(\sigma(fc_1(a_{ptc})) \cdot fc_2(a_{dih})\right) \quad (7.2)$$

where a_{ptc} is a vector of flattened activations from the bottleneck layer of the PTC, a_{dih} is a vector of activations from the last convolutional layer of the DIH encoder, fc_n are trainable fully-connected layers with 512 neurons each, and σ is a softmax activation.

In both the PTC+DIH and PTC+att+DIH, the encoder of the PTC is frozen during training, as in Chapter 6, however in the case of PTC+DIH, the decoder of the PTC is allowed to learn, while in the PTC+att+DIH only the encoder is used. The PTC does not receive any additional supervisory signals, such as ground truth object masks, or scene segmentation, only the end-to-end MSE harmonisation loss.

Baselines

The performance of both end-to-end models is evaluated against two baselines - the vanilla DIH (without semantic segmentation branch), which requires input masks (*DIH-M*), and a no-mask version of the same model (*DIH-NM*), where masks are not provided as input during training. To ensure a fair comparison, all models (bar the frozen part of the PTC) are trained from random initialisation, using the iHarmony dataset and evaluated on the COCO-Exp dataset from Section 7.3.3 and the iHarmony validation set. This is motivated by the fact that the original PTC implementation is only conditioned on exposure shifts, so a comparison across both datasets can illustrate the performance for simple exposure shifts (COCO-Exp) versus more complex colour transformations (iHarmony). If the perceptually-based features learned by the PTC generalise well across image features,

an improvement should be seen over the naive DIH-NM model when evaluated on both these datasets, as indicated by the qualitative evaluation in Chapter 6.

7.6.2 Optimization Details

All models are trained for 50 epochs using the entire training set of the iHarmony dataset, consisting of 65742 training images and evaluated using the validation set, consisting of 7404 validation images. The Adam optimizer (Kingma and Ba, 2014) with default parameters and an initial learning rate of 0.001 is used. The batch size is set to 32 and a 256×256 resolution is enforced. Pre-processing is applied to all input images, namely the pixel intensity range is scaled from $[0, 255]$ to $[-1.0, 1.0]$. For each training run, the model weights corresponding to the lowest validation loss are selected for further evaluation.

7.6.3 Evaluation

In order to compare the performance of the proposed model against the baseline models, the similarity metrics introduced in Section 7.3.5 are used, namely MSE (Eq. 7.1) and PSNR (Eq. 2.3). These metrics are calculated between the ground truth images and the harmonisation results, for each of the models under evaluation.

7.7 End-to-End Model: Results

This section presents the evaluation of the proposed models on both the validation set of the iHarmony dataset, as well as the COCO-Exp dataset used in the two-stage evaluation, introduced in Section 7.3.3.

Table 7.2 shows average MSE and PSNR values for both datasets and each of the models. Both of the proposed end-to-end models improve performance on both the iHarmony and COCO-Exp datasets, as compared to the naive baseline, when performing harmonization with no input mask. This suggests the PTC features are relevant to the image harmonisation task. Overall, the PTC+DIH achieves the best performance in harmonisation with no input mask, outperforming the PTC+att+DIH and the DIH-NM baseline.

Figure 7.9 illustrates the performance of all models under evaluation for several images from the COCO-Exp dataset. Specifically, in each row the input and ground truth are shown in Figures 7.9a and 7.9b respectively. Figures 7.9c, 7.9e and 7.9g show the harmonised outputs of the DIH-NM, PTC+att+DIH and PTC+DIH models respectively, while Figures 7.9d, 7.9f and 7.9h are difference image heatmaps between the input and the harmonised output predicted by each model. These heatmaps provide an illustration of the magnitude, direction and location of the applied correction. Upon inspection of similarity metrics, the harmonised outputs and the difference heatmaps, it can be seen that the PTC+DIH model outperforms both the baseline (DIH-NM) and the latent-space-based combination of both models (PTC+att+DIH). This can be seen clearly when comparing

Model	iHarmony		COCO-Exp	
	MSE	PSNR	MSE	PSNR
DIH-M	89	32.56	201	32.18
DIH-NM	153	30.93	276	31.12
PTC+att+DIH	151	31.02	264	31.37
PTC+DIH	124	31.39	214	31.61

Table 7.2: Test metrics for all evaluated models, across the two datasets used in our experiments. Lower is better for MSE, higher is better for PSNR. Best results using no input mask in bold. Results for the input-mask-based baseline (DIH-M) shown for reference. Higher is better for PSNR, lower is better for MSE.

the difference images: the PTC+DIH applies corrections more consistently across the region of the target object, compared to the two alternatives. Figure 7.10 compares the performance of the PTC+DIH to the mask-based DIH-M model for 3 versions of an input image from iHarmony. It can be noticed that the output of both the PTC+DIH and DIH-M closely follow that of the reference. The area corrected by the PTC+DIH aligns with the ground truth mask. Small differences in the output images can be noted, particularly around edges, where the PTC+DIH sometimes contributes to softness and smearing (e.g. Fig.7.10e, middle row). This is often related to artefacts around the edges of objects and near edges of images produced by the PTC. Nonetheless, despite the lack of input mask, the PTC+DIH achieves consistent and comparable results for each of the image variations and, in some cases, avoids the colour shifts induced by the DIH (e.g. compare columns d) and e) with column c) of Figure 7.10), as discussed in Section 7.5.

Examples of failure cases can be seen in Figure 7.11. The top two rows illustrate the most common failure case, where the region requiring harmonisation is not detected, and thus not corrected by the model. The top row illustrates this scenario for a larger object size, while the middle row does so for a small object (one of the sheep near the bottom of the image). The bottom row shows a scenario where the harmonisation is performed on the correct object, however the amount of correction is insufficient. In addition, the model applies harmonisation to a part of the image not requiring harmonisation (the screen). This behaviour is likely due to the fact that the PTC was originally conditioned on exposure shifts, resulting in higher sensitivity to over-exposure, compared to other image distortions.

The impact of object size on harmonisation performance of all models is summarised in Table 7.3 for both the iHarmony and COCO-Exp datasets. Because the MSE is calculated across the entire image, errors are overall lower for smaller objects. However, when comparing the MSE of harmonised images against their baseline MSE (calculated between the input image and ground truth), the relative MSE improvements are greatest for larger objects. This trend is present across both datasets. The PTC+DIH achieves the lowest errors in each object size category across both datasets. Notably, for objects

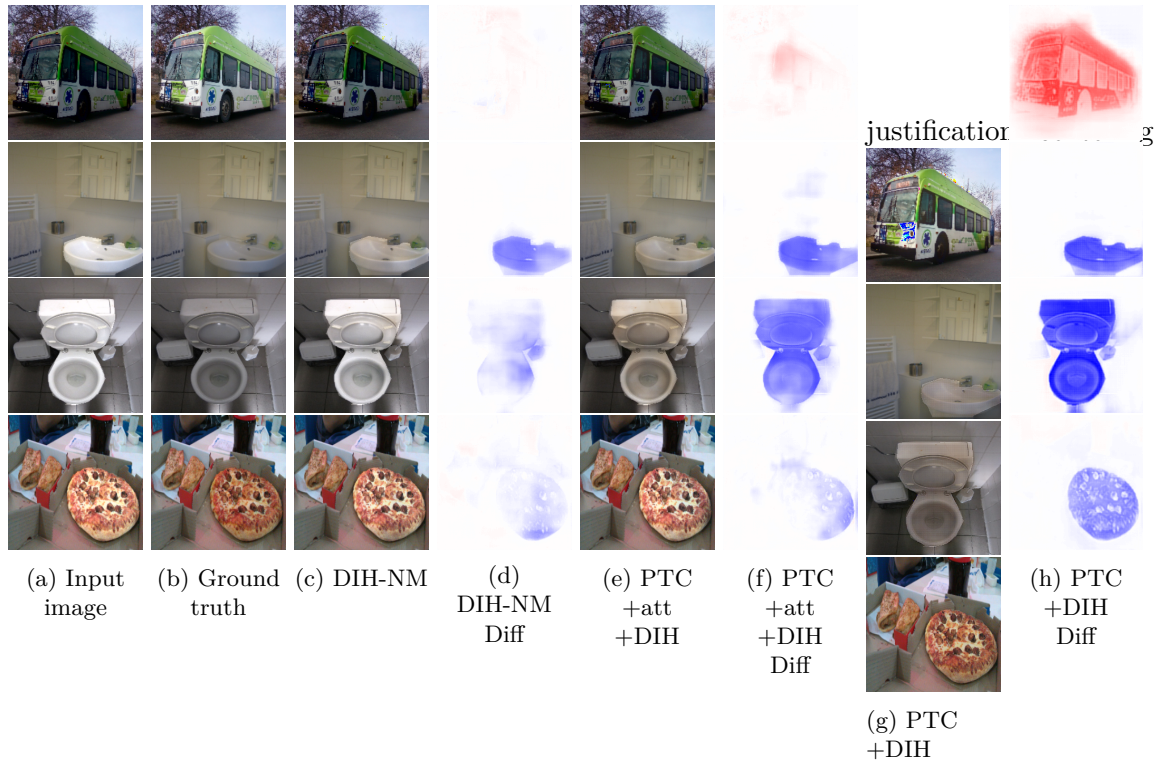


Figure 7.9: Comparison of outputs from each model under evaluation for a range of images from the COCO-Exp dataset. *a)* input image *b)* ground truth *c)* DIH-NM result *d)* Difference image between input and output for DIH-NM *e)* PTC+att+DIH result *f)* difference image between input and output for PTC+att+DIH *g)* PTC+DIH result *h)* PTC+DIH difference image. In difference heatmap images, red indicates a positive difference (i.e. harmonised region is brighter than the corresponding input region), blue indicates the opposite.

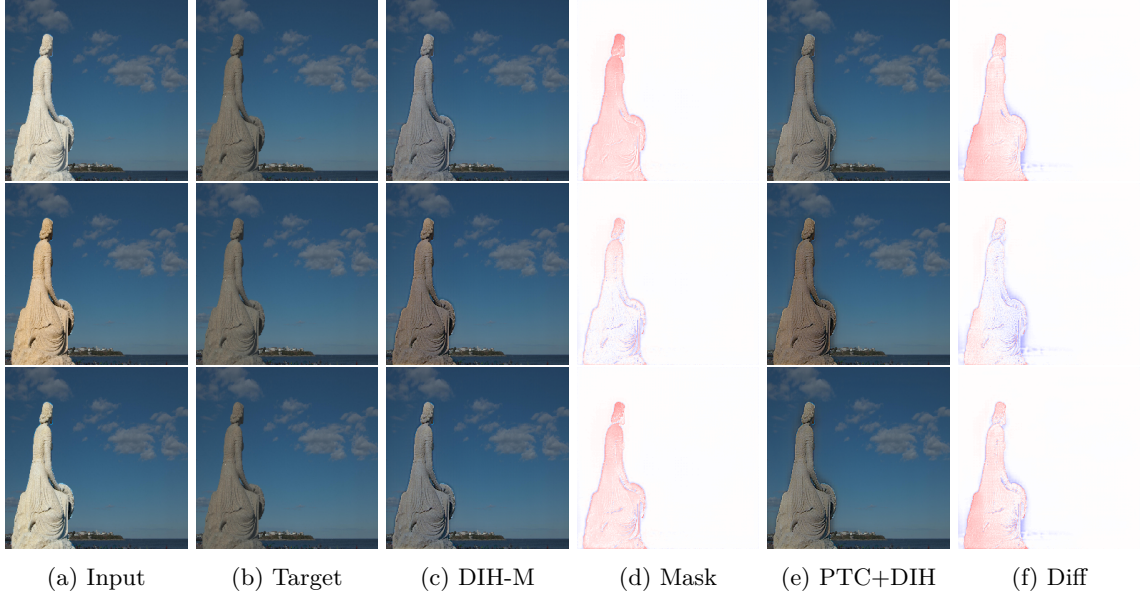


Figure 7.10: Comparison between the corrections applied by PTC+DIH, and the mask-based DIH-M models for multiple variants of the same image. a) input composite, b) ground truth image c) output of DIH-M, d) Difference heatmap between output of DIH-M and ground truth, e) output of PTC+DIH, f) Difference heatmap between output of PTC+DIH and ground truth. In difference heatmap images, red indicates a positive difference (i.e. the harmonised region is brighter than the corresponding input region), blue indicates the opposite.

in the COCO-Exp dataset with areas ranging 20-40% of the image size, the PTC+DIH model achieves lower errors than the mask-based DIH-M baseline. This illustrates the impact of the PTC being conditioned on only exposure shifts, but also indicates that these features are useful when transferred to a different type of transformations, such as those in iHarmony. The performance of the proposed model is also illustrated for images from the authentic composite image dataset proposed by Xue et al. (2012) and mentioned in previous chapters. Examples of this can be seen in Figure 7.12.

7.8 Discussion

The results of both experiments indicate that, in the context of image harmonisation, perceptually-based detection of harmonisation targets can be used to remove the requirement for input object masks. While the proposed approach does not outperform baseline mask-based approaches, it performs significantly better than the state-of-the-art baseline when trained with no input masks. Furthermore, despite the PTC being only conditioned on exposure shifts, its combination with the DIH model improves results on both datasets, suggesting that the perceptually-based features learned by the PTC are useful to the harmonisation task. This is reinforced by the fact that even combining PTC and DIH features in latent space affords a modest improvement over the baseline. Some bias towards exposure shifts is nonetheless noticeable - the largest improvements across both datasets occur for achromatic objects (e.g. the sink or toilet in Fig. 7.9). This could be addressed by training the PTC on a wider range of local transformations. The problem

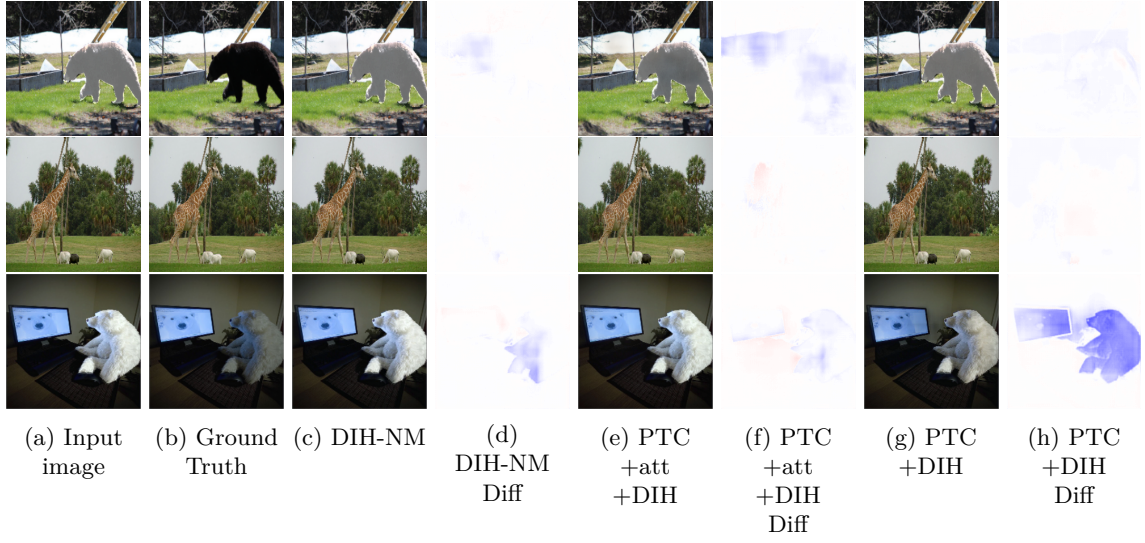


Figure 7.11: Examples of failure cases. *a)* input image *b)* ground truth *c)* DIH-NM result *d)* Difference image between input and output for DIH-NM *e)* PTC+att+DIH result *f)* difference image between input and output for PTC+att+DIH *g)* PTC+DIH result *h)* PTC+DIH difference image. In difference heatmap images, red indicates a positive difference (i.e. the harmonised region is brighter than the corresponding input region), blue indicates the opposite.

iHarmony								
Object Size	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
DIH-M	33.0	116.1	206.5	335.05	456.2	485.48	484.58	705.12
MSE orig.	47.1	235.02	449.84	642.75	1170.31	1222.97	1151.83	1752.12
DIH-NM	50.73	192.22	360.98	497.42	919.29	1058.39	888.11	1534.94
PTC+att+DIH	50.36	190.2	370.65	462.72	884.22	1001.85	933.02	1659.24
PTC+DIH	45.02	150.04	311.72	359.99	623.03	895.33	720.82	1464.62

COCO-Exp								
Object Size	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
DIH-M	73.74	401.55	655.11	785.35	927.68	1042.68	1119.19	1129.01
MSE orig	86.11	524.29	878.42	1131.53	1503.27	1802.57	2072.08	2097.13
DIH-NM	94.3	502.63	828.69	1045.05	1373.97	1661.55	1876.75	1958.01
PTC+att+DIH	93.26	492.65	802.24	986.49	1271.15	1510.16	1684.99	1806.24
PTC+DIH	82.35	410.08	647.13	778.76	946.99	1084.28	1240.54	1295.38

Table 7.3: Average MSE on the iHarmony and COCO-Exp datasets for each of the evaluated models, grouped by area of harmonised object as a fraction of image size. *MSE orig* is the MSE between unharmonised inputs and ground truth. Bold values indicate the lowest error for each object size, given no mask input. DIH-M model shown for reference.



Figure 7.12: Examples of authentic composite images from Xue et al. (2012) processed using the *DIH* model. Leftmost images illustrate the input, middle images illustrate harmonised output, whereas rightmost images show normalised absolute difference image, indicating image locations harmonised by the model. No input masks are supplied.

of object size and its impact on harmonisation accuracy is likely connected to the fact that larger objects tend to contribute to the MSE more, compared to smaller objects. The MSE for a small object requiring a 0.5 stop exposure shift will be lower than that of a larger object requiring the same shift. To alleviate this, when training with input masks, the MSE can simply be scaled by the mask size Sofiuk, Popenova and Konushin (2020), however with no input mask, estimation of target object area becomes non-trivial and presents an interesting direction for further research.

Not unlike the original DIH implementation, the proposed end-to-end model can suffer from gradient artefacts along mask edges, particularly when the initial error to be corrected is large. This issue could be addressed by adopting masked convolutions and utilising self-attention mechanisms, as in Cong et al. (2020) or by explicitly incorporating gradient information, as in Wu et al. (2019). While these issues will be addressed in future work, the advantages of the proposed model demonstrated in this work still hold in the context of image harmonisation with no input mask. Following arguments from Chapter 6, the results confirm that in order to improve image harmonisation performance, particularly in scenarios where input masks are not available, detection of target regions for harmonisation should leverage intermediate representations, equivariant to the transformations of the input to be harmonised. Input masks used in state-of-the-art harmonisation algorithms mimic this role - they encode the presence and location of all input transformations requiring harmonisation as a local binary feature, thus receiving a form of an extra supervisory signal. The results presented here show that explicitly incorporating the artefact detection paradigm into the harmonisation process can be beneficial, while alleviating the requirements for presence of object masks at inference time.

7.9 Conclusions

This chapter has presented and evaluated a novel method for performing image harmonisation without the need for input object masks. The proposed approach leverages two state-of-the-art models - an artefact detector and a harmoniser - which, when combined, produce competitive results to mask-based models. A two-stage evaluation of the original pre-trained models is first performed, and based on evaluation results, this is extended to a custom end-to-end model in two variants, trained from scratch on the challenging iHarmony dataset. Both variants of the proposed end-to-end model are found to outperform the baselines, when evaluated on two different datasets. These findings indicate that information about location and magnitude of composite artefacts can be useful in improving the performance of existing compositing and harmonisation approaches. This is motivated by illustrating that ground truth object masks commonly used in harmonisation algorithms essentially substitute the process of detecting local transformations and inconsistencies requiring correction. Accordingly, the results show that the requirement for provision of object masks for such algorithms can be relaxed or removed entirely by the explicit combination of composite artefact detection with their

correction. This provides a basis for investigation in future work of joint modelling of both the detection and correction of composite image artefacts, e.g. under a multitask learning paradigm, where a joint latent representation is conditioned both to be equivariant with respect to input transformations and to encode the structure of the image. In such a scenario, input masks may be used during the training stage, but would not be necessary during inference.

Chapter 8

Conclusion

8.1 Overview

This thesis has investigated modelling of human perceptual sensitivity in the context of improvement of image composite realism. This was accomplished based on work in three key stages.

First, in Chapters 2 and 4, a formulation of image composite artefacts as local image transformations was developed, and psychometric methods were adopted to model subjective observer sensitivity to these transformations. This was complemented in Chapter 5 by an investigation into the spatial and temporal allocation of overt visual attention by observers performing composite realism assessment in the presence of different transformation types.

Second, to investigate the generalisability of these psychometric models, in Chapter 6 deep learning techniques (introduced and reviewed in Chapter 3) were adopted to learn feature representations of input images equivariant to local exposure transformations. Using transfer learning techniques, these models were then adapted to classify pixels of input images, based on human perceptual characteristics, represented by the proposed empirical psychometric models.

Finally, the trained models were evaluated in a task representing the overarching goal of this thesis: composite image harmonisation. In these experiments, the proposed models produced improved results over baselines, indicating the benefit of the techniques proposed in this work.

As a whole, this thesis presents a novel and practical framework for perceptual modelling of JNDs as a function of local transformations, or distortions, applied to natural images. By adopting synthetic data generation methods and learning-based techniques, empirical psychometric models generated for a relatively small dataset can be extended to novel image content, without the need for collection of additional perceptual data. Importantly,

the deep learning models developed using this framework remove the need for explicit indication of target objects in harmonisation models and outperform state-of-the-art baselines.

The following sections detail and discuss the findings and contributions of this thesis, as well as outlining potential limitations and future work. For clarity, these are discussed on a chapter-by-chapter basis.

8.2 Summary of Findings

8.2.1 Modelling Human Assessment of Image Composite Realism

Defining Image Composite Realism

Chapter 2 discussed fundamental properties of human vision in the context of detection of visual inconsistencies, or distortions. Specifically, many similarities between the problem of image composite realism and image quality assessment were highlighted, placing both in the wider context of subjective visual property modelling. Visual realism in the context of image composites was found to fit the concept of *photorealism* (Ferwerda, 2003). According to this definition, an image is considered photorealistic when it elicits the same visual response as the corresponding real scene. Deviations from this state of affairs result in a decrease in photorealism, provided observers can detect them. Both subjective image quality and photorealism were found to incorporate detection of particular visual properties or inconsistencies and their mapping to a subjective score.

A review of prior work indicated that subjective evaluation methods have been effective at modelling subjective visual properties with a high degree of accuracy. For example, psychometric methods have been successfully used to model subjective realism as a function of a particular image feature, such as shadow softness (Rademacher et al., 2001) or local transformations (Xue et al., 2012). However, this is at the cost of significant effort required to measure subjective human responses under appropriately controlled experimental conditions, which makes generalisation of these models to new image content challenging.

Objective methods, such as the error sensitivity framework (Wang, Bovik and Lu, 2002), structural similarity framework (Wang et al., 2004), natural image statistics (Sheikh, Bovik and Cormack, 2005), or various perceptual visual quality metrics such as the VDP Daly (1992) addressed some of these issues. They were shown to be practical approaches to incorporating knowledge about the HVS into reusable algorithms, which could be applied to a broad range of visual stimuli. This was accomplished through extraction and weighted pooling of different image features, calibrated to correlate with subjective human opinion. However, these approaches were often shown to be complex to design and expensive to adapt to novel tasks or features, as well as not transferring well to complex, real-world stimuli, such as natural images.

Consequently, this chapter defined assessment of visual realism in the context of image composites as detection of visible local inconsistencies and their subsequent pooling into a visual realism rating. While the set of possible inconsistencies was found to be vast, it was also shown to follow particular trends when considering differences between objects and scene in image composites. These include statistical differences in the distributions of low level image properties, such as luminance, colour or contrast (Xue et al., 2012), as well as physical properties of the scene, such as the directionality (Ostrovsky, Cavanagh and Sinha, 2005) and colour (Xue et al., 2012) of the illumination, surface properties (Pardo, Suero and Pérez, 2018), reflections (Cavanagh, Chao and Wang, 2008) and semantics (Biederman, 1981).

Based on this information, psychometric techniques for the modelling of visual realism as a function of local image transformations were developed in subsequent chapters.

Empirical Modelling of Composite Realism

Chapter 4 proposed a novel methodology and empirical models of composite realism, based on using JNDs to define thresholds at which transformations applied to natural images become visible to an average observer. This was achieved by measuring observer sensitivity to local image transformations under a 2AFC paradigm. Observers were required to distinguish between original versions of an image and ones with local transformations applied to a known object. This sensitivity was then measured across a range of 11 transformation magnitudes, 165 scenes and 75 observers and expressed as group JNDs. In order to allow for controlled measurements, synthetic image composites were developed by applying local exposure, contrast and correlated colour temperature transformations to object regions in natural images. In this context, using authentic image composites, for which no realistic ground truth version were readily available, would have not allowed for accurate measurement of JNDs.

In line with Wichmann and Hill (2001a), sigmoidal functions were found to provide a good fit to both group and individual realism ratings as a function of transformation magnitude. Subsequently, group JNDs were extracted by pooling responses for a particular transformation magnitude across observers and scenes. A review of the parameters of the fit models indicated that exposure and CCT transformations were detected most reliably by observers, while contrast transformations, particularly its reduction, yielded higher lapse rates, indicating a larger proportion of observers could not reliably distinguish between transformed and original images. Qualitative evaluation of these group JNDs, accomplished by visualising them as transformations applied to objects in natural images, showed that context and allocation of visual attention may play a significant role in how visible such transformations are to the average observer. Furthermore, transformations resulting in plausible results or applied to semantically ambiguous objects often do not get readily detected, indicating further impact of context. Additionally, as shown previously by Xue et al. (2012), observers consistently displayed a degree of tolerance to certain

transformations, such as contrast or CCT shifts, particularly at low magnitudes. Increased tolerance to exposure and contrast shifts applied to objects near visible illumination sources was also observed in the qualitative evaluation, providing further indication of nonlinear influence of local context.

Overall, this chapter illustrated that the proposed methodology to modelling average group JNDs based on error visibility and transformation magnitude is viable, resulting in well-fit empirical models and interpretable perceptual threshold measures. Equally, several additional questions were raised by the obtained results, particularly concerning the impact of visual attention, either due to task, scene content, or observer knowledge about object identity, on the resulting JNDs.

Allocation of Visual Attention

Chapter 5 undertook further investigation into how the realism assessment task is performed by observers. Specifically, a gaze-based study focusing on the spatial allocation of visual attention was carried out. A 2×2 factorial design was used, evaluating the impact of transformation feature type (exposure and CCT) and prior knowledge of object identity on gaze metrics and resulting subjective realism ratings across 4 groups of observers.

The results illustrated that observers relied primarily on the target object region when assessing realism, regardless of whether prior knowledge regarding the identity of the object was provided. This highlighted the importance of the appearance of the object and its immediate surroundings to the subjective ratings of its realism. Qualitative analysis of fixation maps also illustrated their similarity to saliency, or conspicuity maps. The relative direction of differences in gaze metrics between groups with and without prior object knowledge persisted across the transformation type condition, suggesting that for both CCT and exposure, the provision of prior object knowledge to observers results in comparable changes in visual behaviour. Specifically, a reduction in visual search and scene analysis, and an increase in focus on the target object. This was illustrated by relative decreases in total fixation counts, and increases in their duration. Also, overall response times were found to be shorter when observers were aware of the identity of the object. Interestingly, no significant realism rating differences were found between observer groups with and without prior object knowledge. This is despite the existence of significant differences in gaze metrics between the groups, suggesting that visual search plays an important role in locating composite artefacts, but not necessarily their subjective rating. Similarly, in line with Yarbus (1967), task-related differences in gaze metrics were observed. Specifically, significant differences in the proportion of fixations on target objects were found between presentations of reference and test stimuli.

Despite being highly correlated, the difference in realism ratings across the local transformation type factor was more pronounced compared to the prior object knowledge factor. This confirms prior findings that different image features and distortions may

require individual perceptual models, e.g. as implemented in the work of Xue et al. (2012).

At a high level, this chapter explored and illustrated the impact of image composite distortions on the spatial and temporal allocation of visual attention. It showed that under the experimental conditions outlined, allocation of visual attention and prior knowledge of the object does not impact subjective realism ratings significantly. Additionally, this chapter illustrated that human fixation maps in this task resemble object-centric saliency, or conspicuity maps - indicating the location of transformations, or distortions. This finding was crucially exploited in Chapter 6 in order to model observer responses.

8.2.2 Learning the Observer Function

Machine Learning for Perceptual Modelling

Chapter 3 focused on introducing the background on fundamental machine learning concepts, as well as discussing recent work in learning-based function approximation using deep convolutional neural networks. A number of existing approaches relying on deep learning to approximate perceptual functions were then reviewed, highlighting the architectures, optimisation strategies and datasets used to train such models. Specifically, fully convolutional, image-to-image networks were found suitable for approximating perceptually-informed functions such as saliency prediction (Zhao et al., 2015; Li and Yu, 2015), similarity metrics (Zhang et al., 2018b), image quality (Bosse et al., 2017) or scene understanding (Long, Shelhamer and Darrell, 2015). This research illustrated that such perceptual functions, mapping input images to representations of subjective visual properties, can be approximated from empirical data and generalised to novel image content.

Transformation Equivariant Representations for Perceptual Modelling

Building on findings from Chapters 2-5, Chapter 6 developed a novel methodology for approximation of the realism assessment function performed by observers. This was accomplished by first modelling JNDs with respect to synthetic composites affected by local exposure transformations using the experimental methodology developed in Chapter 4. However, in this study, the JNDs were measured separately for each scene, as opposed to being averaged across scenes, in order for the impact of scene context to be encoded in the subjective responses.

Together with the analysis of visual attention carried out in Chapter 5, this allowed for the formulation of the problem as a supervised learning task - mapping from a synthetic composite image to an output transformation map describing the presence of local exposure transformations above the perceptual thresholds defined by the JNDs. This allowed for the application of convolutional neural network techniques discussed in Chapter 3 to the problem. However, due to the relatively small size of the JND dataset, transfer

learning techniques were required to first learn a feature representation equivariant to the transformations present in the input composites, before fine-tuning the model on the empirical JND data. This was accomplished by adapting a technique proposed by Zhang et al. (2019), which learns to predict the parameter of transformations applied to an input image. This approach was extended to predict not only the parameter of the transformation, but also its location in the image, thus allowing for prediction of local transformations.

Consequently, a self-supervised approach could be adopted to first learn this feature representation. This was done by sampling transformation magnitudes from a distribution, applying them to images and training a model to regress the local parameter of the transformation, given both the original and transformed images as input. After training to convergence, the encoder of this model was then extracted and used as a feature extractor in a pixel-wise classifier fine-tuned on the JND dataset to predict whether each pixel of an input image contains the effects of suprathreshold exposure transformations. The resulting model was found to predict image-wise JNDs with an average error of 0.11 exposure stops. These findings illustrated that, provided a representation equivariant with respect to a specific transformation can be learned, using a small perceptually-conditioned training set is sufficient to approximate the function performed by an observer, or group thereof, performing a local transformation detection task. As many distortions affecting composite images can be defined in the context of transformations, this method can be expanded to a range of other transformation types. Finally, the model was shown to localise composited objects in authentic, rather than synthetic, composite images, prompting evaluation in a wider application context.

8.2.3 Application to Harmonisation

Chapter 7 presented an application of the model developed in Chapter 6 to the problem of composite image harmonisation. This was accomplished by combining the final model trained on JNDs, with a state-of-the-art image harmonisation network. In this arrangement, the model from Chapter 6 served as a perceptual detector - indicating areas requiring harmonisation, based on human perception and removing the need for specifying input object masks. After an initial two-stage evaluation performed on pre-trained, off-the-shelf versions of the models, a full end-to-end model was designed and trained from scratch, using the iHarmony training set (Cong et al., 2020). The proposed novel no-reference end-to-end harmonisation model was found to outperform no-reference versions of baseline models across two harmonisation datasets, indicating the benefit of utilising perceptually-conditioned features for this task. To the author’s knowledge, this is the first end-to-end harmonisation system explicitly modelling perceptually-informed detection of composite artefacts, as well as their harmonisation.

8.2.4 Broader Impact

Viewed as a whole, this thesis has proposed a general framework for modelling visual realism of image composites as a function of local transformations approximating common composite artefacts. Through reliance on signal detection theory and psychometric methods, this framework can be easily adapted to detect different types of relative transformations in natural images, in alignment with subjective perception of realism. The impact of such transformations can also be modelled with respect to subjective properties other than visual realism, provided they are appropriately defined to observers participating in experiments.

At a fundamental level, this was accomplished by formulating the subjective realism assessment process as a function performed by observers, which takes an image composite as input and outputs a map describing the local magnitude of some perceptual property. This function can then be measured directly for a relatively small set of images and approximated directly in the image domain using self-supervised and transfer learning techniques. This allows for practical application of such models to novel stimuli, enabling broader application. This was evidenced by application of the framework discussed in this thesis to the development of a deep learning model, capable of improving performance over baselines in the perceptually-informed task of image harmonisation.

Another benefit of the proposed methodology is its modular nature - depending on the application, the psychometric JND models, training data, neural network architectures or optimisation strategies can be updated. For example, the same transformation equivariant representation can be used as a basis for fine-tuning on JNDs of naive and expert viewer groups. This is possible because the self-supervised training process of this representation does not involve any human-perceptual bias - this is only introduced in the fine-tuning stage. This is visualised in Chapter 1, Figure 1.5, which provides an overview of the framework developed in this work.

8.3 Limitations and Future Work

While the models proposed in this thesis have been shown to improve image harmonisation results, several considerations and limitations should be addressed in future work.

8.3.1 Data Collection and Perceptual Models

While the proposed methodology reduces the need for collection of empirical data, it does not remove it completely. Consequently, development of empirical perceptual models remains a effort-intensive element of the proposed framework. Future work should focus on assessing the impact of the size and characteristics of the perceptual training dataset on the accuracy of resulting models. This could allow for a reduction in the number of images for which JNDs need to be measured. Another consideration relates to the impact of the

type of transformations the model should detect and the corresponding requirement for training data. It can be expected that for complex or subtle transformations, adjustments to the perceptual data collection and modelling process may have to be made depending on average observer performance. As such, more subjective data collection would likely be necessary before training this model to detect novel transformations. The benefit of the presented approach, however, is that much of the computationally-expensive training of the detection models can be accomplished without subjective perceptual data, in the pre-training stage. Similarly, expansion of the range of transformation parameters the model is sensitive to, is only bound by the perceptual fine-tuning stage. Here, more perceptual data would need to be collected to cover the wider range of parameters, in order to update the perceptual model. The pre-training stage can be adapted by simply adjusting the data generation parameters. In practice, this decoupling of feature learning and perceptual fine-tuning thus affords flexibility when applying this model to other related tasks.

8.3.2 Model Optimisation, Scalability and Transformation Types

In order to allow for controlled evaluation of the proposed approach, the models developed in this thesis focus on a select few transformation types. While this was a necessary constraint, it potentially limits wider applications of proposed models, without additional fine-tuning. As the transformation equivariant representation learning task does not require perceptually labelled data, future work should focus on training such representations with respect to a much wider range of local transformations and evaluating the impact of such conditioning on performance in auxiliary tasks. Since training labels can be generated automatically for this part of the process, improving the learned representation would require arguably less effort compared to collection of new perceptual data.

With frequent advances in the field of DL architectures, potential incremental improvements to the accuracy and efficiency of the proposed models could be made by utilising novel architectures. For example, weakly-supervised methods, such as few-shot learning, provide alternative approaches for scenarios where training data is very limited. Throughout the process of writing of this thesis, many novel, general-purpose deep learning models have achieved state-of-the-art results on challenging datasets. While it was outside of the scope of this work to perform a large scale evaluation, this remains an interesting direction for future work, and the modular design of the proposed methodology certainly supports this.

Finally, the efficiency of the deep learning models developed in this thesis could benefit from application of model distillation techniques (Polino, Pascanu and Alistarh, 2018), allowing for a reduction in the number of model parameters. While the large number of parameters can be beneficial for training on large, synthetically-generated datasets, it has a negative impact on training and inference times, resulting in longer iteration times during experimentation. Any future work, attempting to extend this model, should consider first

applying distillation techniques.

8.3.3 Application to Other Tasks & Alternative Approaches

The experiments presented in Chapter 7 illustrated that the proposed models can improve state-of-the-art harmonisation results. However, potential improvements to generalisability could be afforded by focusing efforts on the transformation equivariance of the representation with respect to a wide array of local transformations. As suggested by Bengio, Goodfellow and Courville (2017) and confirmed in Chapter 6, this property is fundamental to accurate modelling of human perception in this context, allowing for separation of features of the same image affected by different transformations in feature space. Provided these properties are well encoded, sub- and supra-threshold classification could be accomplished by using linear classifiers. Ultimately, such a feature representation should learn to encode effects of transformations in natural images. Rather than attempting to classify specific types of transformations, it should describe statistical deviations from distributions of pixels found in natural images. Thus, the more general this feature representation is with respect to encoding various local transformations, the simpler the task of perceptual tuning should become. However, in order to achieve this, future work should address scalable, self-supervised methods of learning such representations, while attempting to minimise the number of required perceptually-annotated training examples. In addition to this, several novel approaches to this problem could be adopted. This includes leveraging pre-trained discriminator networks from large-scale GANs for image synthesis, such as the BigGAN (Brock, Donahue and Simonyan, 2018), and leveraging their deep features as descriptors, following the methodology described by Zhang et al. (2018b). Furthermore, recent developments in computational cognitive neuroscience (Mishra and Majhi, 2019) and the increasing interplay with deep learning (Kietzmann, McClure and Kriegeskorte, 2019) provide interesting, alternative approaches to the problems tackled in this thesis. Development of biologically-plausible computational models of human vision, not unlike those described by Marr (1982), would enable a more fundamental understanding of the impact of various image distortions. Such knowledge would largely simplify the parameterisation of perceptual models, as well as allow training of deep networks with biologically-inspired, rather than statistically-approximated, perceptual losses.

8.4 Final Comments

This thesis has investigated the use of machine learning techniques to learn and generalise psychometric models of subjective perception of visual realism in composite images. This was achieved by first using psychometric frameworks in order to empirically model human perception of realism in synthetic image composites, in the presence of controlled local image transformations. Both this experimental methodology and the models were then validated in the context of deployment of visual attention, before being employed to

collect a set of image-wise empirical perceptual thresholds, used as training data for the machine learning models. The following work focused on developing techniques to train convolutional neural networks to detect local exposure transformations in a self-supervised manner, before fine-tuning the resulting transformation-equivariant representation on the empirical perceptual threshold data. The resulting model was then applied to an image composite harmonisation task and compared against existing approaches, achieving comparable results and removing the requirement for object segmentation data to be provided to the network at inference time. Finally, aside from the datasets and developed models, this work proposed a methodological framework for learning to automatically detect local artefacts in images, based on generalisation of psychometric models through leveraging data synthesis and machine learning techniques.

Bibliography

- Adelson, E.H., Bergen, J.R. et al., 1991. *The plenoptic function and the elements of early vision*, vol. 2. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of . . .
- Ahumeda, A., 1996. Simplified vision models for image-quality assessment. *Sid international symposium digest of technical papers*. Society for Information Display, vol. 27, pp.397–402.
- Akcay, S., Kundegorski, M.E., Willcocks, C.G. and Breckon, T.P., 2018. Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. *IEEE transactions on information forensics and security*, 13(9), pp.2203–2215.
- Antes, J.R., 1974. The time course of picture viewing. *Journal of experimental psychology*, 103(1), p.62.
- Arevalo, J., González, F.A., Ramos-Pollán, R., Oliveira, J.L. and Lopez, M.A.G., 2016. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods and programs in biomedicine*, 127, pp.248–257.
- Arregui, X.Q., Geday, M.A., García, M.C., Otón, E. and Otón, J.M., 2020. Image sensors for digital photography: a short course for undergraduates. i: Optics. *Optica pura y aplicada*, 53(1), p.5.
- Assembly, C.X.P., 1990. Report 1082-1, studies toward the unification of picture assessment methodology. *Reports of the CCIR, XI-Part*, 1, pp.384–414.
- Azadi, S., Pathak, D., Ebrahimi, S. and Darrell, T., 2018. Compositional gan: Learning conditional image composition. *arXiv preprint arXiv:1807.07560*.
- Bak, C., Kocak, A., Erdem, E. and Erdem, A., 2017. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, 20(7), pp.1688–1698.
- Banterle, F., Ledda, P., Debattista, K., Bloj, M., Artusi, A. and Chalmers, A., 2009. A psychophysical evaluation of inverse tone mapping techniques. *Computer graphics forum*. Wiley Online Library, vol. 28, pp.13–25.
- Barbour, C.G. and Meyer, G.W., 1992. Visual cues and pictorial limitations for computer generated photorealistic images. *The Visual Computer*, 9(3), pp.151–165.
- Barrow, H. and Tenenbaum, J., 1978. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2.
- Barten, P.G., 1999. *Contrast sensitivity of the human eye and its effects on image quality*, vol. 19. Spie optical engineering press Bellingham, WA.
- Bengio, Y., Courville, A. and Vincent, P., 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), pp.1798–1828.
- Bengio, Y., Goodfellow, I. and Courville, A., 2017. *Deep learning*, vol. 1. Citeseer.

- Bengio, Y., Louradour, J., Collobert, R. and Weston, J., 2009. Curriculum learning. *Proceedings of the 26th annual international conference on machine learning*. pp.41–48.
- Bergstra, J. and Bengio, Y., 2012. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb), pp.281–305.
- Bertalmio, M., Sapiro, G., Caselles, V. and Ballester, C., 2000. Image inpainting. *Proceedings of the 27th annual conference on computer graphics and interactive techniques*. pp.417–424.
- Bianco, S., Celona, L., Napoletano, P. and Schettini, R., 2018. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2), pp.355–362.
- Biederman, I., 1972. Perceiving real-world scenes. *Science*, 177(4043), pp.77–80.
- Biederman, I., 1981. *On the semantics of a glance at a scene*.
- Biederman, I., Mezzanotte, R.J. and Rabinowitz, J.C., 1982. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2), pp.143–177.
- Birch, J., 1997. Efficiency of the ishihara test for identifying red-green colour deficiency. *Ophthalmic and Physiological Optics*, 17(5), pp.403–408.
- Borbély, Á., Sámson, Á. and Schanda, J., 2001. The concept of correlated colour temperature revisited. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(6), pp.450–457.
- Borji, A., 2019. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*.
- Borji, A., Sihite, D.N. and Itti, L., 2013. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1), pp.55–69.
- Bosse, S., Maniry, D., Müller, K.R., Wiegand, T. and Samek, W., 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1), pp.206–219.
- Boureau, Y.L., Ponce, J. and LeCun, Y., 2010. A theoretical analysis of feature pooling in visual recognition. *Proceedings of the 27th international conference on machine learning (icml-10)*. pp.111–118.
- Boyce, S.J. and Pollatsek, A., 1992. Identification of objects in scenes: the role of scene background in object naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3), p.531.
- Brock, A., Donahue, J. and Simonyan, K., 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. and Shah, R., 1994. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*. pp.737–744.
- Bubic, A., Von Cramon, D.Y. and Schubotz, R.I., 2010. Prediction, cognition and the brain. *Frontiers in human neuroscience*, 4, p.25.
- Bylinskii, Z., Borkin, M.A., Kim, N.W., Pfister, H. and Oliva, A., 2015. Eye fixation metrics for large scale evaluation and comparison of information visualizations. *Workshop on eye tracking and visualization*. Springer, pp.235–255.

- Campbell, F.W. and Robson, J., 1968. Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3), pp.551–566.
- Caruana, R., 1997. Multitask learning. *Machine learning*, 28(1), pp.41–75.
- Cauchy, A., 1847. Méthode générale pour la résolution des systemes d’équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847), pp.536–538.
- Cavanagh, P., 2005. The artist as neuroscientist. *Nature*, 434(7031), p.301.
- Cavanagh, P., Chao, J. and Wang, D., 2008. Reflections in art. *Spatial vision*, 21(3), pp.261–270.
- Caviedes, J. and Gurbuz, S., 2002. No-reference sharpness metric based on local edge kurtosis. *Image processing. 2002. proceedings. 2002 international conference on.* IEEE, vol. 3, pp.III–III.
- Chen, B.C. and Kae, A., 2019. Toward realistic image compositing with adversarial learning. *Proceedings of the ieee conference on computer vision and pattern recognition.* pp.8415–8424.
- Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S. and Urtasun, R., 2016. Monocular 3d object detection for autonomous driving. *Proceedings of the ieee conference on computer vision and pattern recognition.* pp.2147–2156.
- Cheng, Z., Yang, Q. and Sheng, B., 2015. Deep colorization. *Proceedings of the ieee international conference on computer vision.* pp.415–423.
- Choi, M.G., Jung, J.H. and Jeon, J.W., 2009. No-reference image quality assessment using blur and noise. *International Journal of Computer Science and Engineering*, 3(2), pp.76–80.
- Chopra, S., Hadsell, R. and LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. *2005 ieee computer society conference on computer vision and pattern recognition (cvpr’05).* IEEE, vol. 1, pp.539–546.
- Chou, C.H. and Li, Y.C., 1995. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Transactions on circuits and systems for video technology*, 5(6), pp.467–476.
- Chou, C.H. and Liu, K.C., 2008. Colour image compression based on the measure of just noticeable colour difference. *IET Image Processing*, 2(6), pp.304–322.
- Clark, A., 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3), pp.181–204.
- Clevert, D.A., Unterthiner, T. and Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.
- Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W. and Zhang, L., 2020. Dovenet: Deep image harmonization via domain verification. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition.* pp.8394–8403.
- Connors, R.W. and Ng, C.T., 1989. Developing a quantitative model of human preattentive vision. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6), pp.1384–1407.
- Cornia, M., Baraldi, L., Serra, G. and Cucchiara, R., 2016. A deep multi-level network for saliency prediction. *2016 23rd international conference on pattern recognition (icpr).* IEEE, pp.3488–3493.
- Corriveau, P., Gojmerac, C., Hughes, B. and Stelmach, L., 1999. All subjective scales are not created equal: The effects of context on different scales. *Signal processing*, 77(1), pp.1–9.

- Cox, D.R., 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), pp.215–232.
- Cranor, L.F., 2008. A framework for reasoning about the human in the loop.
- Cui, J., Wen, F. and Tang, X., 2008. Real time google and live image search re-ranking. *Proceedings of the 16th acm international conference on multimedia*. pp.729–732.
- Cun, X. and Pun, C.M., 2020. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29, pp.4759–4771.
- Curtin, M. and Vanderhoef, J., 2015. A vanishing piece of the pi: The globalization of visual effects labor. *Television & New Media*, 16(3), pp.219–239.
- Dalal, N. and Triggs, B., 2005. Histograms of oriented gradients for human detection. *2005 IEEE computer society conference on computer vision and pattern recognition (cvpr'05)*. IEEE, vol. 1, pp.886–893.
- Daly, S.J., 1992. Visible differences predictor: an algorithm for the assessment of image fidelity. *Human vision, visual processing, and digital display iii*. International Society for Optics and Photonics, vol. 1666, pp.2–16.
- Desimone, R. and Duncan, J., 1995. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1), pp.193–222.
- Doersch, C., Gupta, A. and Efros, A.A., 2015. Unsupervised visual representation learning by context prediction. *Proceedings of the IEEE international conference on computer vision*. pp.1422–1430.
- Dolhasz, A., Frutos-Pascual, M. and Williams, I., 2017. Composite Realism: Effects of Object Knowledge and Mismatched Feature Type on Observer Gaze and Subjective Quality. *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE, pp.9–14.
- Dolhasz, A., Harvey, C. and Williams, I., 2020. Learning to Observe: Approximating Human Perceptual Thresholds for Detection of Suprathreshold Image Transformations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp.4797–4807.
- Dolhasz., A., Harvey., C. and Williams., I., 2020. Towards unsupervised image harmonisation. *Proceedings of the 15th international joint conference on computer vision, imaging and computer graphics theory and applications - volume 5: Visapp.*. INSTICC, SciTePress, pp.574–581. Available from: <http://doi.org/10.5220/0009354705740581>.
- Dolhasz, A., Williams, I. and Frutos-Pascual, M., 2016. Measuring Observer Response to Object-Scene Disparity in Composites. *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE, pp.13–18.
- Doukakis, E., Debattista, K., Bashford-Rogers, T., Dhokia, A., Asadipour, A., Chalmers, A. and Harvey, C., 2019. Audio-visual-olfactory resource allocation for tri-modal virtual environments. *IEEE transactions on visualization and computer graphics*, 25(5), pp.1865–1875.
- Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. and Pal, C., 2016. The importance of skip connections in biomedical image segmentation. *Deep learning and data labeling for medical applications*, Springer, pp.179–187.
- Duchowski, A.T., 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, 34(4), pp.455–470.
- Eagleman, D.M., 2001. Visual illusions and neurobiology. *Nature Reviews Neuroscience*, 2(12), pp.920–926.

- Eckersley, P. and Nasser, Y., 2017. Eff ai progress measurement project. Available from: <https://eff.org/ai/metrics>.
- Efron, B. and Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pp.54–75.
- Efron, B. and Tibshirani, R.J., 1994. *An introduction to the bootstrap*. CRC press.
- Eigen, D. and Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *Proceedings of the ieee international conference on computer vision*. pp.2650–2658.
- Einhäuser, W. and König, P., 2003. Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, 17(5), pp.1089–1097.
- Elhelw, M., Nicolaou, M., Chung, A., Yang, G.Z. and Atkins, M.S., 2008. A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Transactions on Applied Perception (TAP)*, 5(1), p.3.
- Emberson, L.L., 2019. How does learning and memory shape perceptual development in infancy? *Psychology of learning and motivation*, Elsevier, vol. 70, pp.129–160.
- Engelke, U., Pepion, R., Le Callet, P. and Zepernick, H.J., 2010. Linking distortion perception and visual saliency in h. 264/avc coded video containing packet loss. *Visual communications and image processing 2010*. International Society for Optics and Photonics, pp.774406–774406.
- Eskicioglu, A.M. and Fisher, P.S., 1995. Image quality measures and their performance. *IEEE Transactions on communications*, 43(12), pp.2959–2965.
- Evgeniou, T. and Pontil, M., 2004. Regularized multi-task learning. *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining*. ACM, pp.109–117.
- Fan, S., Ng, T.T., Herberg, J.S., Koenig, B.L., Tan, C.Y.C. and Wang, R., 2014. An automated estimator of image visual realism based on human cognition. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.4201–4208.
- Fan, S., Ng, T.T., Koenig, B.L., Herberg, J.S., Jiang, M., Shen, Z. and Zhao, Q., 2018. Image visual realism: From human perception to machine computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(9), pp.2180–2193.
- Fechner, G., 1966. Elements of psychophysics. vol. i. German original published in 1860.
- Ferreira, W.D., Ferreira, C.B., Cruz Júnior, G. da and Soares, F., 2020. A review of digital image forensics. *Computers & Electrical Engineering*, 85, p.106685.
- Ferwerda, J.A., 2003. Three varieties of realism in computer graphics. *Human vision and electronic imaging viii*. International Society for Optics and Photonics, vol. 5007, pp.290–298.
- Ferzli, R. and Karam, L.J., 2009. A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). *IEEE transactions on image processing*, 18(4), pp.717–728.
- Fiedler, M., Hossfeld, T. and Tran-Gia, P., 2010. A generic quantitative relationship between quality of experience and quality of service. *IEEE Network*, 24(2).
- Fleming, R.W., Dror, R.O. and Adelson, E.H., 2003. Real-world illumination and the perception of surface reflectance properties. *Journal of vision*, 3(5), pp.3–3.

- Fleming, R.W., Torralba, A. and Adelson, E.H., 2004. Specular reflections and the perception of shape. *Journal of vision*, 4(9), pp.10–10.
- Fodor, J.A., Pylyshyn, Z.W. et al., 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), pp.3–71.
- Freeman, W.T., 1994. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368(6471), p.542.
- Friedman, A., 1979. Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of experimental psychology: General*, 108(3), p.316.
- Fründ, I., Haenel, N.V. and Wichmann, F.A., 2011. Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, 11(6). Available from: <http://doi.org/10.1167/11.6.16>.
- Gai, E.G. and Curry, R.E., 1976. A model of the human observer in failure detection tasks. *IEEE Transactions on Systems, Man, and Cybernetics*, (2), pp.85–94.
- Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C. and Lalonde, J.F., 2017. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*.
- Gatys, L.A., Ecker, A.S. and Bethge, M., 2016. Image style transfer using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp.2414–2423.
- Gegenfurtner, A., Lehtinen, E. and Säljö, R., 2011. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, 23(4), pp.523–552.
- Geisler, W.S., 2011. Contributions of ideal observer theory to vision research. *Vision research*, 51(7), pp.771–781.
- Gibson, J.J., 1950. The perception of visual surfaces. *The American journal of psychology*, 63(3), pp.367–384.
- Gibson, J.J., 1966. The senses considered as perceptual systems.
- Gidaris, S., Singh, P. and Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.
- Girshick, R., 2015. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*. pp.1440–1448.
- Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp.580–587.
- Glassner, A.S., 2014. *Principles of digital image synthesis*. Elsevier.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. *Advances in neural information processing systems*. pp.2672–2680.
- Gordo, A., Almazán, J., Revaud, J. and Larlus, D., 2016. Deep image retrieval: Learning global representations for image search. *European conference on computer vision*. Springer, pp.241–257.
- Gregory, R.L., 1970. The intelligent eye.

- Gu, K., Zhai, G., Yang, X. and Zhang, W., 2014. Deep learning network for blind image quality assessment. *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp.511–515.
- Guillemot, C. and Le Meur, O., 2013. Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1), pp.127–144.
- Hans-Hermann, B., 2008. Origins and extensions of the k-means algorithm in cluster analysis. *Journal Electronique d’Histoire des Probabilités et de la Statistique Electronic Journal for History of Probability and Statistics*, 4(2).
- Harel, J., Koch, C. and Perona, P., 2007. Graph-based visual saliency. *Advances in neural information processing systems*. pp.545–552.
- He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp.770–778.
- Helmholtz, H.v., 1856. Treatise of physiological optics: Concerning the perceptions in general. *Classics in psychology*, pp.79–127.
- Henderson, J.M. and Hollingworth, A., 1998. Eye movements during scene viewing: An overview. *Eye guidance in reading and scene perception*, 11, pp.269–293.
- Henderson, J.M. and Hollingworth, A., 1999. High-level scene perception. *Annual review of psychology*, 50(1), pp.243–271.
- Hill, N.J., 2001. *Testing hypotheses about psychometric functions*. Ph.D. thesis. University of Oxford.
- Hinton, G.E., Krizhevsky, A. and Wang, S.D., 2011. Transforming auto-encoders. *International conference on artificial neural networks*. Springer, pp.44–51.
- Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02), pp.107–116.
- Hohwy, J., 2013. *The predictive mind*. Oxford University Press.
- Holladay, J.T., 2004. Visual acuity measurements. *Journal of Cataract & Refractive Surgery*, 30(2), pp.287–290.
- Horn, R.A., 1990. The hadamard product. *Proc. symp. appl. math.* vol. 40, pp.87–169.
- Hornik, K., Stinchcombe, M., White, H. et al., 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), pp.359–366.
- Hosu, V., Lin, H., Sziranyi, T. and Saupe, D., 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29, pp.4041–4056.
- Hou, W., Gao, X., Tao, D. and Li, X., 2014. Blind image quality assessment via deep learning. *IEEE transactions on neural networks and learning systems*, 26(6), pp.1275–1286.
- Hurvich, L.M. and Jameson, D., 1957. An opponent-process theory of color vision. *Psychological review*, 64(6p1), p.384.
- Huynh-Thu, Q., Garcia, M.N., Speranza, F., Corriveau, P. and Raake, A., 2011. Study of rating scales for subjective quality assessment of high-definition video. *IEEE Transactions on Broadcasting*, 57(1), pp.1–14.
- Intraub, H., 2002. Visual scene perception. *Encyclopedia of cognitive science*, 4, pp.524–527.

- Ioffe, S. and Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.
- Ishihara, S. et al., 1918. Tests for color blindness. *American Journal of Ophthalmology*, 1(5), p.376.
- Isola, P., Zhu, J.Y., Zhou, T. and Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp.1125–1134.
- Itti, L. and Koch, C., 2001. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3), p.194.
- Itti, L., Koch, C. and Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11), pp.1254–1259.
- ITU, I., 2002. 500-11, “methodology for the subjective assessment of the quality of television pictures,” recommendation itu-r bt. 500-11. *ITU Telecom. Standardization Sector of ITU*, 7.
- ITU-R BT, R., 2002. 500-11, methodology for the subjective assessment of the quality of television pictures”. *International Telecommunication Union, Tech. Rep.*
- Jacob, R. and Karn, K.S., 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, 2(3), p.4.
- Janssen, T. and Blommaert, F., 1997. Image quality semantics. *Journal of imaging science and Technology*, 41(5), pp.555–560.
- Jia, Y., Lin, W. and Kassim, A.A., 2006. Estimating just-noticeable distortion for video. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(7), pp.820–829.
- Jiang, L., Xu, M., Liu, T., Qiao, M. and Wang, Z., 2018. Deepvs: A deep learning based video saliency prediction approach. *Proceedings of the European conference on computer vision (eccv)*. pp.602–617.
- Johnson, J., Alahi, A. and Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. *European conference on computer vision*. Springer, pp.694–711.
- Jones, B.L. and McManus, P.R., 1986. Graphic scaling of qualitative terms. *SMPTE journal*, 95(11), pp.1166–1171.
- Jumisko, S.H., Ilvonen, V.P. and Vaananen-Vainio-Mattila, K.A., 2005. Effect of tv content in subjective assessment of video quality on mobile devices. *Multimedia on mobile devices*. International Society for Optics and Photonics, vol. 5684, pp.243–254.
- Kajiya, J.T., 1986. The rendering equation. *Proceedings of the 13th annual conference on computer graphics and interactive techniques*. pp.143–150.
- Kane, C.L., 2011. Satiated and denied: War and visual realism. *Afterimage*, 39(1/2), p.46.
- Kang, L., Ye, P., Li, Y. and Doermann, D., 2015. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. *2015 IEEE international conference on image processing (ICIP)*. IEEE, pp.2791–2795.
- Keeffe, J.E., Lovie-Kitchin, J.E., Maclean, H. and Taylor, H.R., 1996. A simplified screening test for identifying people with low vision in developing countries. *Bulletin of the World Health Organization*, 74(5), p.525.
- Kersten, D., Mamassian, P. and Yuille, A., 2004. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55, pp.271–304.

- Kietzmann, T.C., McClure, P. and Kriegeskorte, N., 2019. Deep neural networks in computational neuroscience. *Oxford research encyclopedia of neuroscience*.
- Kim, J. and Lee, S., 2017. Deep learning of human visual sensitivity in image quality assessment framework. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp.1676–1684.
- Kim, Y.J., Luo, M.R., Choe, W., Kim, H.S., Park, S.O., Baek, Y., Rhodes, P., Lee, S. and Kim, C.Y., 2008. Factors affecting the psychophysical image quality evaluation of mobile phone displays: the case of transmissive liquid-crystal displays. *JOSA A*, 25(9), pp.2215–2222.
- Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Klambauer, G., Unterthiner, T., Mayr, A. and Hochreiter, S., 2017. Self-normalizing neural networks. *Advances in neural information processing systems*. pp.971–980.
- Kleffner, D.A. and Ramachandran, V.S., 1992. On the perception of shape from shading. *Perception & Psychophysics*, 52(1), pp.18–36.
- Klein, A., Falkner, S., Bartels, S., Hennig, P. and Hutter, F., 2016. Fast bayesian optimization of machine learning hyperparameters on large datasets. *arXiv preprint arXiv:1605.07079*.
- Klein, G. and Murray, D.W., 2009. Simulating low-cost cameras for augmented reality compositing. *IEEE transactions on visualization and computer graphics*, 16(3), pp.369–380.
- Knill, D.C. and Richards, W., 1996. *Perception as bayesian inference*. Cambridge University Press.
- Koch, C. and Ullman, S., 1987. Shifts in selective visual attention: towards the underlying neural circuitry. *Matters of intelligence*, Springer, pp.115–141.
- Koenderink, J.J., Doorn, A.J. van and Pont, S.C., 2004. Light direction from shaded random gaussian surfaces. *Perception*, 33(12), pp.1405–1420.
- Kohler, J., Daneshmand, H., Lucchi, A., Zhou, M., Neymeyr, K. and Hofmann, T., 2018. Towards a theoretical understanding of batch normalization. *stat*, 1050, p.27.
- Komogortsev, O.V., Gobert, D.V., Jayarathna, S., Gowda, S.M. et al., 2010. Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, 57(11), pp.2635–2645.
- Kong, S., Shen, X., Lin, Z., Mech, R. and Fowlkes, C., 2016. Photo aesthetics ranking network with attributes and content adaptation. *European conference on computer vision*. Springer, pp.662–679.
- König, S.D. and Buffalo, E.A., 2014. A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds. *Journal of neuroscience methods*, 227, pp.121–131.
- Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. pp.1097–1105.
- Krogh, A. and Hertz, J.A., 1992. A simple weight decay can improve generalization. *Advances in neural information processing systems*. pp.950–957.
- Króliczak, G., Heard, P., Goodale, M.A. and Gregory, R.L., 2006. Dissociation of perception and action unmasked by the hollow-face illusion. *Brain research*, 1080(1), pp.9–16.
- LaBerge, D., 1998. Attentional emphasis in visual orienting and resolving. *Visual attention*, pp.417–454.

- Lalonde, J.F. and Efros, A.A., 2007. Using color compatibility for assessing image realism. *Computer vision, 2007. iccv 2007. IEEE 11th international conference on*. IEEE, pp.1–8.
- Lassalle, J., Gros, L., Morineau, T. and Coppin, G., 2012. Impact of the content on subjective evaluation of audiovisual quality: What dimensions influence our perception? *Ieee international symposium on broadband multimedia systems and broadcasting*. IEEE, pp.1–6.
- Le Meur, O. and Baccino, T., 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1), pp.251–266.
- Le Meur, O., Le Callet, P., Barba, D. and Thoreau, D., 2006. A coherent computational approach to model bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 28(5), pp.802–817.
- Le Meur, O., Ninassi, A., Le Callet, P. and Barba, D., 2010. Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing: Image Communication*, 25(7), pp.547–558.
- Lee, C., Rincon, G.A., Meyer, G., Höllerer, T. and Bowman, D.A., 2013. The effects of visual realism on search tasks in mixed reality simulation. *IEEE transactions on visualization and computer graphics*, 19(4), pp.547–556.
- Legge, G.E. and Foley, J.M., 1980. Contrast masking in human vision. *Josa*, 70(12), pp.1458–1471.
- Leisti, T., Radun, J., Virtanen, T., Halonen, R. and Nyman, G., 2009. Subjective experience of image quality: attributes, definitions, and decision making of subjective image quality. *Image quality and system performance vi*. International Society for Optics and Photonics, vol. 7242, p.72420D.
- Leveque, L., Zhang, W. and Liu, H., 2019. Subjective assessment of image quality induced saliency variation. *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp.1024–1028.
- Levy-Schoen, A., 2017. Flexible and/or rigid control of oculomotor scanning behavior. *Eye movements*, Routledge, pp.299–314.
- Li, G. and Yu, Y., 2015. Visual saliency based on multiscale deep features. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp.5455–5463.
- Li, G. and Yu, Y., 2018. Contrast-oriented deep neural networks for salient object detection. *IEEE transactions on neural networks and learning systems*, 29(12), pp.6038–6051.
- Li, Y. and Pizlo, Z., 2011. Depth cues versus the simplicity principle in 3d shape perception. *Topics in cognitive science*, 3(4), pp.667–685.
- Liechty, J., Pieters, R. and Wedel, M., 2003. Global and local covert visual attention: Evidence from a bayesian hidden markov model. *Psychometrika*, 68(4), pp.519–541.
- Lin, J.P. and Sun, M.T., 2018. A yolo-based traffic counting system. *2018 conference on technologies and applications of artificial intelligence (taai)*. IEEE, pp.82–85.
- Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*. pp.2980–2988.
- Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014. Microsoft coco: Common objects in context. *European conference on computer vision*. Springer, pp.740–755.

- Lin, W. and Kuo, C.C.J., 2011. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4), pp.297–312.
- Liu, B. and Todd, J.T., 2004. Perceptual biases in the interpretation of 3d shape from shading. *Vision research*, 44(18), pp.2135–2145.
- Liu, H. and Heynderickx, I., 2011. Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(7), pp.971–982.
- Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X. and Shum, H.Y., 2011. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2), pp.353–367.
- Loftus, G.R. and Mackworth, N.H., 1978. Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human perception and performance*, 4(4), p.565.
- Lokka, I.E., Çöltekin, A., Wiener, J., Fabrikant, S.I. and Röcke, C., 2018. Virtual environments as memory training devices in navigational tasks for older adults. *Scientific reports*, 8.
- Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp.3431–3440.
- Longuet-Higgins, H.C., 1981. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828), p.133.
- Longuet-Higgins, H.C., Prazdny, K. et al., 1980. The interpretation of a moving retinal image. *Proc. r. soc. lond. b. The Royal Society*, vol. 208, pp.385–397.
- Lopez-Moreno, J., Sundstedt, V., Sangorin, F. and Gutierrez, D., 2010. Measuring the perception of light inconsistencies. *Proceedings of the 7th symposium on applied perception in graphics and visualization*. ACM, pp.25–32.
- Losada, M. and Mullen, K., 1994. The spatial tuning of chromatic mechanisms identified by simultaneous masking. *Vision research*, 34(3), pp.331–341.
- Loshchilov, I. and Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.
- Lowe, D.G., 1999. Object recognition from local scale-invariant features. *Proceedings of the seventh IEEE international conference on computer vision*. Ieee, vol. 2, pp.1150–1157.
- Lubin, J. and Fibush, D., 1997. Sarnoff jnd vision model.
- Lupyan, G. and Clark, A., 2015. Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, 24(4), pp.279–284.
- Ma, S., Liu, J. and Wen Chen, C., 2017. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp.4535–4544.
- Maas, A.L., Hannun, A.Y. and Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. *Proc. icml*. vol. 30, p.3.
- Mannos, J. and Sakrison, D., 1974. The effects of a visual fidelity criterion of the encoding of images. *IEEE transactions on Information Theory*, 20(4), pp.525–536.
- Manocha, P., Finkelstein, A., Jin, Z., Bryan, N.J., Zhang, R. and Mysore, G.J., 2020. A differentiable perceptual audio metric learned from just noticeable differences. *arXiv preprint arXiv:2001.04460*.

- Mantiuk, R., Daly, S.J., Myszkowski, K. and Seidel, H.P., 2005. Predicting visible differences in high dynamic range images: model and its calibration. *Human vision and electronic imaging x*. International Society for Optics and Photonics, vol. 5666, pp.204–214.
- Mantiuk, R., Kim, K.J., Rempel, A.G. and Heidrich, W., 2011. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4), pp.1–14.
- Mantiuk, R.K., Tomaszewska, A. and Mantiuk, R., 2012. Comparison of four subjective methods for image quality assessment. *Computer graphics forum*. Wiley Online Library, vol. 31, pp.2478–2491.
- Marat, S., Phuoc, T.H., Granjon, L., Guyader, N., Pellerin, D. and Guérin-Dugué, A., 2009. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International journal of computer vision*, 82(3), p.231.
- Marius’ t Hart, B., Schmidt, H.C.E.F., Klein-Harmeyer, I. and Einhäuser, W., 2013. Attention in natural scenes: contrast affects rapid visual processing and fixations alike. *Phil. Trans. R. Soc. B*, 368(1628), p.20130067.
- Marius’ t Hart, B., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., Koenig, P. and Einhäuser, W., 2009. Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition*, 17(6-7), pp.1132–1158.
- Marr, D., 1982. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. Inc., New York, NY, 2(4.2).
- Marschner, S.R. and Greenberg, D.P., 1998. *Inverse rendering for computer graphics*. Citeseer.
- Marziliano, P., Dufaux, F., Winkler, S. and Ebrahimi, T., 2002. A no-reference perceptual blur metric. *Image processing. 2002. proceedings. 2002 international conference on*. IEEE, vol. 3, pp.III–III.
- McNamara, A., Chalmers, A., Troscianko, T. and Gilchrist, I., 2000. Comparing real & synthetic scenes using human judgements of lightness. *Rendering techniques 2000*, Springer, pp.207–218.
- McNamara, A. et al., 2005. Exploring perceptual equivalence between real and simulated imagery. *Proceedings of the 2nd symposium on applied perception in graphics and visualization*. ACM, pp.123–128.
- Meijer, F., Geudeke, B.L. and Broek, E.L. Van den, 2009. Navigating through virtual environments: Visual realism improves spatial cognition. *CyberPsychology & Behavior*, 12(5), pp.517–521.
- Mirza, M. and Osindero, S., 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Mishra, A. and Majhi, S.K., 2019. A comprehensive survey of recent developments in neuronal communication and computational neuroscience. *Journal of Industrial Information Integration*, 13, pp.40–54.
- Mitchell, T.M. et al., 1997. Machine learning.
- Moffitt, K., 1980. Evaluation of the fixation duration in visual search. *Perception & Psychophysics*, 27(4), pp.370–372.
- Morgenstern, Y., Murray, R.F. and Harris, L.R., 2011. The human visual system’s assumption that light comes from above is weak. *Proceedings of the National Academy of Sciences*, 108(30), pp.12551–12553.
- Mori, M., MacDorman, K.F. and Kageki, N., 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2), pp.98–100.

- Mukherjee, R., Debattista, K., Bashford-Rogers, T., Vangorp, P., Mantiuk, R., Bessa, M., Waterfield, B. and Chalmers, A., 2016. Objective and subjective evaluation of high dynamic range video compression. *Signal Processing: Image Communication*, 47, pp.426–437.
- Mullen, K.T., 1985. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of physiology*, 359(1), pp.381–400.
- Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R. and Kim, K., 2018. Image to image translation for domain adaptation. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.4500–4509.
- Nair, V. and Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th international conference on machine learning (icml-10)*. pp.807–814.
- Nattkemper, D. and Prinz, W., 1984. Costs and benefits of redundancy in visual search. *Advances in psychology*, Elsevier, vol. 22, pp.343–351.
- Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A. and Yosinski, J., 2017. Plug & play generative networks: Conditional iterative generation of images in latent space. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.4467–4477.
- Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q. and Shen, D., 2017. Medical image synthesis with context-aware generative adversarial networks. *International conference on medical image computing and computer-assisted intervention*. Springer, pp.417–425.
- Ninassi, A., Le Meur, O., Le Callet, P. and Barba, D., 2007. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. *2007 ieee international conference on image processing*. IEEE, vol. 2, pp.II–169.
- Ninassi, A., Le Meur, O., Le Callet, P., Barba, D. and Tirel, A., 2006. Task impact on the visual attention in subjective image quality assessment. *Signal processing conference, 2006 14th european*. IEEE, pp.1–5.
- Noble, W.S., 2006. What is a support vector machine? *Nature biotechnology*, 24(12), pp.1565–1567.
- Noroozi, M. and Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. *European conference on computer vision*. Springer, pp.69–84.
- Odén, A. and Wedel, H., 1975. Arguments for fisher’s permutation test. *The Annals of Statistics*, pp.518–520.
- Ojala, T., Pietikainen, M. and Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7), pp.971–987.
- Oliva, A., 2005. Gist of the scene. *Neurobiology of attention*, Elsevier, pp.251–256.
- Oliva, A. and Torralba, A., 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155, pp.23–36.
- Orquin, J.L. and Loose, S.M., 2013. Attention and choice: A review on eye movements in decision making. *Acta psychologica*, 144(1), pp.190–206.
- Ostrovsky, Y., Cavanagh, P. and Sinha, P., 2001. Perceiving illumination inconsistencies in scenes.
- Ostrovsky, Y., Cavanagh, P. and Sinha, P., 2005. Perceiving illumination inconsistencies in scenes. *Perception*, 34(11), pp.1301–1314.

- Owen, A.B., 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2), pp.237–249.
- Palmer, S.E., 1999. *Vision science: Photons to phenomenology*. MIT press.
- Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E. and Nieto, X. Giro-i, 2017. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*.
- Pan, J., Sayrol, E., Nieto, X. Giro-i, McGuinness, K. and O'Connor, N.E., 2016. Shallow and deep convolutional networks for saliency prediction. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp.598–606.
- Pan, S.J. and Yang, Q., 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), pp.1345–1359.
- Pardo, P.J., Suero, M.I. and Pérez, Á.L., 2018. Correlation between perception of color, shadows, and surface textures and the realism of a scene in virtual reality. *JOSA A*, 35(4), pp.B130–B135.
- Park, T., Liu, M.Y., Wang, T.C. and Zhu, J.Y., 2019. Semantic image synthesis with spatially-adaptive normalization. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp.2337–2346.
- Patel, Y., Appalaraju, S. and Manmatha, R., 2019. Human perceptual evaluations for image compression. *arXiv preprint arXiv:1908.04187*.
- Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E. and Lindeløv, J.K., 2019. Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51(1), pp.195–203.
- Perazzi, F., Krähenbühl, P., Pritch, Y. and Hornung, A., 2012. Saliency filters: Contrast based filtering for salient region detection. *Computer vision and pattern recognition (cvpr), 2012 IEEE conference on*. IEEE, pp.733–740.
- Perez, L. and Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.
- Pinson, M.H. and Wolf, S., 2003. Comparing subjective video quality testing methodologies. *Visual communications and image processing 2003*. International Society for Optics and Photonics, vol. 5150, pp.573–583.
- Pizlo, Z., 2014. *Making a machine that sees like us*. Oxford University Press (UK).
- Poggio, T., Torre, V. and Koch, C., 1987. Computational vision and regularization theory. *Readings in computer vision*, Elsevier, pp.638–643.
- Polino, A., Pascanu, R. and Alistarh, D., 2018. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*.
- Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F. et al., 2013. Color image database tid2013: Peculiarities and preliminary results. *European workshop on visual information processing (euwip)*. IEEE, pp.106–111.
- Pont, S.C. and Pas, S.F. te, 2006. Material—illumination ambiguities and the perception of solid objects. *Perception*, 35(10), pp.1331–1350.

- Porter, T. and Duff, T., 1984. Compositing digital images. *Acm siggraph computer graphics*. ACM, vol. 18, pp.253–259.
- Potter, M.C., 1976. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, 2(5), p.509.
- Prince, S., 2010. Through the looking glass: Philosophical toys and digital visual effects. *Projections*, 4(2), pp.19–40.
- Prince, S., 2011. *Digital visual effects in cinema: the seduction of reality*. Rutgers University Press.
- Qi, G.J., Zhang, L., Chen, C.W. and Tian, Q., 2019. Avt: Unsupervised learning of transformation equivariant representations by autoencoding variational transformations. *Proceedings of the ieee international conference on computer vision*. pp.8130–8139.
- Qian, N., 1999. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1), pp.145–151.
- Rademacher, P., Lengyel, J., Cutrell, E. and Whitted, T., 2001. Measuring the perception of visual realism in images. *Rendering techniques 2001*, Springer, pp.235–247.
- Radford, A., Metz, L. and Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Ragan, E.D., Bowman, D.A., Kopper, R., Stinson, C., Scerbo, S. and McMahan, R.P., 2015. Effects of field of view and visual complexity on virtual reality training effectiveness for a visual scanning task. *IEEE transactions on visualization and computer graphics*, 21(7), pp.794–807.
- Rajashekar, U., Cormack, L.K. and Bovik, A.C., 2004. Point-of-gaze analysis reveals visual search strategies. *Electronic imaging 2004*. International Society for Optics and Photonics, pp.296–306.
- Ramachandran, V.S., 1988. Perceiving shape from shading. *Scientific American*, 259(2), pp.76–83.
- Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3), p.372.
- Redi, J., Liu, H., Zunino, R. and Heynderickx, I., 2011. Interactions of visual attention and quality perception. *IsEt/spie electronic imaging*. International Society for Optics and Photonics, pp.78650S–78650S.
- Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.779–788.
- Redmon, J. and Farhadi, A., 2017. Yolo9000: better, faster, stronger. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.7263–7271.
- Redmon, J. and Farhadi, A., 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H., 2016. Generative adversarial text to image synthesis. *International conference on machine learning*. PMLR, pp.1060–1069.
- Reichle, E.D., Pollatsek, A., Fisher, D.L. and Rayner, K., 1998. Toward a model of eye movement control in reading. *Psychological review*, 105(1), p.125.
- Reinhard, E., Ashikhmin, M., Gooch, B. and Shirley, P., 2001. Color transfer between images. *IEEE Computer graphics and applications*, (5), pp.34–41.

- Reinhard, E., Efros, A.A., Kautz, J. and Seidel, H.P., 2013. On visual realism of synthesized imagery. *Proceedings of the IEEE*, 101(9), pp.1998–2007.
- Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*. pp.91–99.
- Ribeiro, F., Florencio, D. and Nascimento, V., 2011. Crowdsourcing subjective image quality evaluation. *2011 18th ieee international conference on image processing*. IEEE, pp.3097–3100.
- Riche, N., Duvinage, M., Mancas, M., Gosselin, B. and Dutoit, T., 2013. Saliency and human fixations: state-of-the-art and study of comparison metrics. *Proceedings of the ieee international conference on computer vision*. pp.1153–1160.
- Robertson, A.R., 1968. Computation of correlated color temperature and distribution temperature. *JOSA*, 58(11), pp.1528–1535.
- Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International conference on medical image computing and computer-assisted intervention*. Springer, pp.234–241.
- Ruder, S., 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.
- Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning representations by back-propagating errors. *nature*, 323(6088), pp.533–536.
- Russell, B.C., Torralba, A., Murphy, K.P. and Freeman, W.T., 2008. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3), pp.157–173.
- Salvucci, D.D. and Goldberg, J.H., 2000. Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the 2000 symposium on eye tracking research & applications*. pp.71–78.
- Sarikaya, D., Corso, J.J. and Guru, K.A., 2017. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. *IEEE transactions on medical imaging*, 36(7), pp.1542–1549.
- Schmidt, U. and Roth, S., 2012. Learning rotation-aware features: From invariant priors to equivariant descriptors. *2012 ieee conference on computer vision and pattern recognition*. IEEE, pp.2050–2057.
- Seshadrinathan, K., Soundararajan, R., Bovik, A.C. and Cormack, L.K., 2010a. Study of subjective and objective quality assessment of video. *IEEE transactions on image processing*, 19(6), pp.1427–1441.
- Seshadrinathan, K., Soundararajan, R., Bovik, A.C. and Cormack, L.K., 2010b. A subjective study to evaluate video quality assessment algorithms. *Human vision and electronic imaging xv*. International Society for Optics and Photonics, vol. 7527, p.75270H.
- Sharafi, Z., Shaffer, T., Sharif, B. and Guéhéneuc, Y.G., 2015. Eye-tracking metrics in software engineering. *2015 asia-pacific software engineering conference (apsec)*. IEEE, pp.96–103.
- Sheikh, H.R. and Bovik, A.C., 2004. Image information and visual quality. *Acoustics, speech, and signal processing, 2004. proceedings.(icassp'04). ieee international conference on*. IEEE, vol. 3, pp.iii–709.
- Sheikh, H.R., Bovik, A.C. and Cormack, L., 2005. No-reference quality assessment using natural scene statistics: Jpeg2000. *IEEE Transactions on image processing*, 14(11), pp.1918–1927.
- Sheikh, H.R., Bovik, A.C. and De Veciana, G., 2005. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12), pp.2117–2128.

- Sheikh, H.R., Sabir, M.F. and Bovik, A.C., 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11), pp.3440–3451.
- Sheikh, H.R., Wang, Z., Cormack, L. and Bovik, A.C., 2005. Live image quality assessment database release 2 (2005).
- Shi, W., Loy, C.C. and Tang, X., 2016. Deep specialized network for illuminant estimation. *European conference on computer vision*. Springer, pp.371–387.
- Simons, D.J. and Levin, D.T., 1997. Change blindness. *Trends in cognitive sciences*, 1(7), pp.261–267.
- Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Smallman, H.S. and John, M.S., 2005. Naïve realism: Misplaced faith in realistic displays. *Ergonomics in design*, 13(3), pp.6–13.
- Smith, A.R., 1995. Alpha and the history of digital compositing. URL: http://www.alvyray.com/Memos/7_alpha.pdf, zuletzt abgerufen am, 24, p.2010.
- Smolensky, P., 1988. On the proper treatment of connectionism. *Behavioral and brain sciences*, 11(1), pp.1–23.
- Snell, J., Swersky, K. and Zemel, R., 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*. pp.4077–4087.
- Snellen, H., 1868. *Probebuchstaben zur bestimmung der sehschärfe*. Verlag von Hermann Peters.
- Socher, R., Ganjoo, M., Manning, C.D. and Ng, A., 2013. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*. pp.935–943.
- Sofiuk, K., Popenova, P. and Konushin, A., 2020. Foreground-aware semantic representations for image harmonization. *arXiv preprint arXiv:2006.00809*.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), pp.1929–1958.
- Stanton, J.M., 2001. Galton, pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9(3).
- Streijl, R.C., Winkler, S. and Hands, D.S., 2016. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2), pp.213–227.
- Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H. and Hospedales, T.M., 2018. Learning to compare: Relation network for few-shot learning. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.1199–1208.
- Sunkavalli, K., Johnson, M.K., Matusik, W. and Pfister, H., 2010. Multi-scale image harmonization. *Acm tog. ACM*, vol. 29, p.125.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.1–9.
- Taigman, Y., Polyak, A. and Wolf, L., 2016. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*.

- Talebi, H. and Milanfar, P., 2018. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8), pp.3998–4011.
- Tan, M., Lalonde, J.F., Sharan, L., Rushmeier, H. and O’sullivan, C., 2015. The perception of lighting inconsistencies in composite outdoor scenes. *ACM Transactions on Applied Perception (TAP)*, 12(4), p.18.
- Teo, P.C. and Heeger, D.J., 1994. Perceptual image distortion. *Image processing, 1994. proceedings. icip-94., ieee international conference*. IEEE, vol. 2, pp.982–986.
- Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C. and Nießner, M., 2016a. Face2face: Real-time face capture and reenactment of rgb videos. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.2387–2395.
- Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C. and Niessner, M., 2016b. Face2face: Real-time face capture and reenactment of rgb videos. *The ieee conference on computer vision and pattern recognition (cvpr)*.
- Tobii, 2012. Tobii x1 light eye tracker. Available from: http://acuity-ets.com/downloads/tobii_x1_eye_tracker_leaflet.pdf.
- Tompson, J., Goroshin, R., Jain, A., LeCun, Y. and Bregler, C., 2015. Efficient object localization using convolutional networks. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.648–656.
- Torgerson, W.S., 1958. Theory and methods of scaling.
- Torralba, A. and Oliva, A., 2003. Statistics of natural image categories. *Network: computation in neural systems*, 14(3), pp.391–412.
- Treisman, A.M. and Gelade, G., 1980. A feature-integration theory of attention. *Cognitive psychology*, 12(1), pp.97–136.
- Treue, S. and Trujillo, J.C.M., 1999. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736), pp.575–579.
- Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X. and Yang, M.H., 2017. Deep image harmonization. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.3789–3797.
- Turing, A., 1950. Computing machinery and intelligence, mind lix, 433-60.
- Turing, A.M., 2009. Computing machinery and intelligence. *Parsing the turing test*, Springer, pp.23–65.
- Van Nes, F.L. and Bouman, M.A., 1967. Spatial modulation transfer in the human eye. *JOSA*, 57(3), pp.401–406.
- Vangorp, P., Laurijssen, J. and Dutré, P., 2007. The influence of shape on the perception of material reflectance. *Acm transactions on graphics (tog)*. ACM, vol. 26, p.77.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*. pp.5998–6008.
- Vecera, S.P. and Rizzo, M., 2003. Spatial attention: normal processes and their breakdown. *Neurologic clinics*, 21(3), pp.575–607.
- Venkatesh, M.V. and Sen-ching, S.C., 2010. Eye tracking based perceptual image inpainting quality analysis. *2010 ieee international conference on image processing*. IEEE, pp.1109–1112.

- Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D. et al., 2016. Matching networks for one shot learning. *Advances in neural information processing systems*. pp.3630–3638.
- Von Helmholtz, H., 1867. *Handbuch der physiologischen optik*, vol. 9. Voss.
- Voronin, V., Franc, V., Zelensky, A. and Agaian, S., 2019. Video quality assessment using generative adversarial network. *Mobile multimedia/image processing, security, and applications 2019*. International Society for Optics and Photonics, vol. 10993, p.109930S.
- Vu, C.T., Larson, E.C. and Chandler, D.M., 2008. Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. *2008 ieee southwest symposium on image analysis and interpretation*. IEEE, pp.73–76.
- Wagberg, J., 2020. optprop - a color properties toolbox (<https://www.mathworks.com/matlabcentral/fileexchange/13788-optprop-a-color-properties-toolbox>), matlab central file exchange. Retrieved April 1, 2020.
- Wagemans, J., Elder, J.H., Kubovy, M., Palmer, S.E., Peterson, M.A., Singh, M. and Heydt, R. von der, 2012. A century of gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. *Psychological bulletin*, 138(6), p.1172.
- Wallis, S.A., Baker, D.H., Meese, T.S. and Georgeson, M.A., 2013. The slope of the psychometric function and non-stationarity of thresholds in spatiotemporal contrast vision. *Vision research*, 76, pp.1–10.
- Walter, B., Pattanaik, S.N. and Greenberg, D.P., 2002. Using perceptual texture masking for efficient image synthesis. *Computer graphics forum*. Wiley Online Library, vol. 21, pp.393–399.
- Wang, G., Li, L., Li, Q., Gu, K., Lu, Z. and Qian, J., 2017. Perceptual evaluation of single-image super-resolution reconstruction. *2017 ieee international conference on image processing (icip)*. IEEE, pp.3145–3149.
- Wang, L., Lu, H., Ruan, X. and Yang, M.H., 2015. Deep networks for saliency detection via local estimation and global search. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.3183–3192.
- Wang, N. and Doube, W., 2011. How real is really? a perceptually motivated system for quantifying visual realism in digital images. *2011 international conference on multimedia and signal processing*. IEEE, pp.141–149.
- Wang, Q., Cavanagh, P. and Green, M., 1994. Familiarity and pop-out in visual search. *Perception & psychophysics*, 56(5), pp.495–500.
- Wang, Q., Xu, L., Chen, Q. and Sun, Q., 2014. Import of distortion on saliency applied to image quality assessment. *2014 ieee international conference on image processing (icip)*. IEEE, pp.1165–1169.
- Wang, S., Zhang, L., Liang, Y. and Pan, Q., 2012. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. *2012 ieee conference on computer vision and pattern recognition*. IEEE, pp.2216–2223.
- Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J. and Catanzaro, B., 2018a. High-resolution image synthesis and semantic manipulation with conditional gans. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.8798–8807.
- Wang, W. and Shen, J., 2017. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5), pp.2368–2378.
- Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y. and Change Loy, C., 2018b. Esrgan:

- Enhanced super-resolution generative adversarial networks. *Proceedings of the european conference on computer vision (eccv)*. pp.0–0.
- Wang, Z. and Bovik, A.C., 2006. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1), pp.1–156.
- Wang, Z. and Bovik, A.C., 2009. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1), pp.98–117.
- Wang, Z., Bovik, A.C. and Lu, L., 2002. Why is image quality assessment so difficult? *2002 ieee international conference on acoustics, speech, and signal processing*. IEEE, vol. 4, pp.IV–3313.
- Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), pp.600–612.
- Wang, Z. and Li, Q., 2010. Information content weighting for perceptual image quality assessment. *IEEE Transactions on image processing*, 20(5), pp.1185–1198.
- Watson, A. and Sasse, M.A., 1996. Evaluating audio and video quality in low-cost multimedia conferencing systems. *Interacting with Computers*, 8(3), pp.255–275.
- Watson, A. and Sasse, M.A., 1998. Measuring perceived quality of speech and video in multimedia conferencing applications. *Proceedings of the sixth acm international conference on multimedia*. ACM, pp.55–60.
- Watson, A.B. and Pelli, D.G., 1983. Quest: A bayesian adaptive psychometric method. *Perception & psychophysics*, 33(2), pp.113–120.
- Weber, E.H., 1996. *Eh weber on the tactile senses*. Psychology Press.
- Weiss, K., Khoshgoftaar, T.M. and Wang, D., 2016. A survey of transfer learning. *Journal of Big data*, 3(1), p.9.
- Welch, R.B., Blackmon, T.T., Liu, A., Mellers, B.A. and Stark, L.W., 1996. The effects of pictorial realism, delay of visual feedback, and observer interactivity on the subjective sense of presence. *Presence: Teleoperators & Virtual Environments*, 5(3), pp.263–273.
- Wichmann, F.A. and Hill, N.J., 2001a. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8), pp.1293–1313.
- Wichmann, F.A. and Hill, N.J., 2001b. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8), pp.1293–1313.
- Wichmann, F.A. and Hill, N.J., 2001c. The psychometric function: Ii. bootstrap-based confidence intervals and sampling. *Perception & psychophysics*, 63(8), pp.1314–1329.
- Wilkening, J. and Fabrikant, S.I., 2011. How do decision time and realism affect map-based decision making? *International conference on spatial information theory*. Springer, pp.1–19.
- Wolfe, J.M., 2000. Visual attention. *Seeing*, 2, pp.335–386.
- Wolpert, D.H. and Macready, W.G., 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1), pp.67–82.
- Wong, B.Y., Shih, K.T., Liang, C.K. and Chen, H.H., 2012. Single image realism assessment and recoloring by color compatibility. *IEEE Transactions on Multimedia*, 14(3), pp.760–769.
- Wright, S., 2013a. *Compositing visual effects: Essentials for the aspiring artist*. Routledge.

- Wright, S., 2013b. *Digital compositing for film and video*. Taylor & Francis.
- Wu, H., Zheng, S., Zhang, J. and Huang, K., 2019. Gp-gan: Towards realistic high-resolution image blending. *Proceedings of the 27th acm international conference on multimedia*. ACM, pp.2487–2495.
- Wu, J., Shi, G. and Lin, W., 2019. Survey of visual just noticeable difference estimation. *Frontiers of Computer Science*, 13(1), pp.4–15.
- Wyszecki, G. and Stiles, W.S., 1982. *Color science*, vol. 8. Wiley New York.
- Xiao, J., Hays, J., Ehinger, K.A., Oliva, A. and Torralba, A., 2010. Sun database: Large-scale scene recognition from abbey to zoo. *Cvpr, 2010 ieee conference on*. IEEE, pp.3485–3492.
- Xie, J., Xu, L. and Chen, E., 2012. Image denoising and inpainting with deep neural networks. *Advances in neural information processing systems*. pp.341–349.
- Xue, S., Agarwala, A., Dorsey, J. and Rushmeier, H., 2012. Understanding and improving the realism of image composites. *ACM Transactions on Graphics (TOG)*, 31(4), p.84.
- Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D. and DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), pp.8619–8624.
- Yang, P., Baracchi, D., Ni, R., Zhao, Y., Argenti, F. and Piva, A., 2020. A survey of deep learning-based source image forensics. *Journal of Imaging*, 6(3), p.9.
- Yang, X., Ling, W., Lu, Z., Ong, E.P. and Yao, S., 2005. Just noticeable distortion model and its applications in video coding. *Signal Processing: Image Communication*, 20(7), pp.662–680.
- Yao, Y., Hu, W., Zhang, W., Wu, T. and Shi, Y.Q., 2018. Distinguishing computer-generated graphics from natural images based on sensor pattern noise and deep learning. *Sensors*, 18(4), p.1296.
- Yarbus, A.L., 1967. *Eye movements during perception of complex objects*. Springer.
- Young, S.R., Rose, D.C., Karnowski, T.P., Lim, S.H. and Patton, R.M., 2015. Optimizing deep learning hyper-parameters through an evolutionary algorithm. *Proceedings of the workshop on machine learning in high-performance computing environments*. pp.1–5.
- Young, T., 1802. Ii. the bakerian lecture. on the theory of light and colours. *Philosophical transactions of the Royal Society of London*, 92, pp.12–48.
- Yuille, A. and Kersten, D., 2006. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7), pp.301–308.
- Zangemeister, W., Sherman, K. and Stark, L., 1995. Evidence for a global scanpath strategy in viewing abstract compared with realistic images. *Neuropsychologia*, 33(8), pp.1009–1025.
- Zeiler, M.D. and Fergus, R., 2014. Visualizing and understanding convolutional networks. *European conference on computer vision*. Springer, pp.818–833.
- Zeiss, V., 2014. Zeiss online vision screening. Available from: <https://www.zeiss.com/vision-care/int/better-vision/zeiss-online-vision-screening-check.html>.
- Zelinsky, G.J., Rao, R.P.N., Hayhoe, M.M. and Ballard, D.H., 1997. Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological science*, 8(6), pp.448–453.
- Zhai, G. and Min, X., 2020. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63, pp.1–52.

- Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. and Metaxas, D.N., 2018a. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8), pp.1947–1962.
- Zhang, L., Qi, G.J., Wang, L. and Luo, J., 2019. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.2547–2555.
- Zhang, R., Isola, P. and Efros, A.A., 2016. Colorful image colorization. *European conference on computer vision*. Springer, pp.649–666.
- Zhang, R., Isola, P., Efros, A.A., Shechtman, E. and Wang, O., 2018b. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.586–595.
- Zhang, R., Isola, P., Efros, A.A., Shechtman, E. and Wang, O., 2018c. The unreasonable effectiveness of deep features as a perceptual metric. *Cvpr*.
- Zhang, Z., Luo, P., Loy, C.C. and Tang, X., 2014. Facial landmark detection by deep multi-task learning. *European conference on computer vision*. Springer, pp.94–108.
- Zhao, R., Ouyang, W., Li, H. and Wang, X., 2015. Saliency detection by multi-context deep learning. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.1265–1274.
- Zhao, Z.Q., Zheng, P., Xu, S.t. and Wu, X., 2019. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11), pp.3212–3232.
- Zhu, J.Y., Krahenbuhl, P., Shechtman, E. and Efros, A.A., 2015. Learning a discriminative model for the perception of realism in composite images. *Proceedings of the ieee iccv*. pp.3943–3951.
- Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the ieee international conference on computer vision*. pp.2223–2232.
- Zhu, X.J., 2005. *Semi-supervised learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

Appendix A

Publications

Measuring Observer Response to Object-Scene Disparity in Composites

Alan Dolhasz*

Ian Williams

Maite Frutos-Pascual

DMT Lab, Birmingham City University

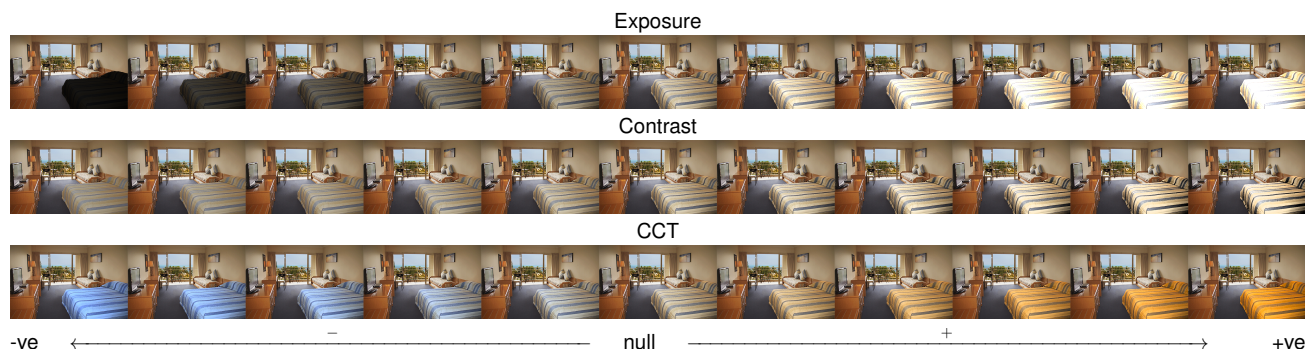


Figure 1: An example of the stimuli used in the experiments. Top row: exposure scaling from 0.1 to 1.9; Middle row: contrast scaling from 0.43 to 2.27; Bottom row: CCT offsets from -200 to +200 mired).

ABSTRACT

Image composites are combinations of image elements from different sources, often combined in a manner to give the appearance of a single, coherent image. This work assesses the impact of low-level image feature offsets on observer response with respect to realism of image composites. The response to selected features, namely exposure, contrast and Correlated Colour Temperature (CCT), is evaluated in a series of 3 experiments, each employing 25 human observers. A total of 10890 data points are analysed, 3630 for each experiment, and psychometric functions are fit to this data in order to parametrise and quantify the relationship between human responses and the amount of disparity between object and scene. Average thresholds and their confidence intervals for each of the image features are then presented and discussed, notably indicating a degree of observer variance in realism responses, particularly in the presence of negative contrast disparities. Exposure, as well as CCT offsets are found to be more readily detected, the latter also contributing to some false positives at high offsets, due to illumination-reflectance ambiguity. The resulting thresholds and confidence intervals can be utilised in creating realistic composites, as well as understanding the impact of different features on observers' perception of composite realism.

Keywords: Composite, realism, subjective quality, object, scene, disparity.

Index Terms: Computing Methodologies [Computer Graphics]: Graphics systems and interfaces—Perception;

1 INTRODUCTION

An image composite is a single image made up of two or more source images, or elements thereof [1]. The aim of compositing is achieving a visually-seamless and coherent combination of the

component image parts - often an object and background scene. Today, composites are used in many image-related fields, from photography [2], through visual effects (VFX) in film [3], to Mixed and Augmented Reality (MR & AR) [4].

In AR, both optical and video see-through head-mounted displays (HMDs) must deal with matching the visual properties of the real world, as it is presented to the viewer, with those of augmentations - here analogous to composites. In the case of video see-through HMDs, the properties of the camera capturing the real world affect the appearance of the image, as seen by the viewer. In the optical see-through scenario, such disparities are even greater, due to the inherent issues with field of view, resolution and contrast [5]. Due to these implicit disparities, it is important to understand the effect they can have on viewers' quality and realism perception. Such issues are notably easier to deal with in traditional, screen-based compositing, where the scene and augmentation share a single display. Here, disparities can be minimised by adjusting a range of low-level image features (for example luminance, contrast or saturation), either in a time-consuming, manual manner [1], or using automated computational approaches [6, 7]. Furthermore, the final quality is influenced by the properties of the object, environment (scene), display device and the individual differences between observers. These issues have been organised and detailed by Kruijff et al. [8].

Composite improvement usually relies on estimation and matching of image-based features [7, 9, 10, 11], recreating and applying scene illumination conditions to the object [12, 13, 14, 15], matching texture-based features [16] and improving the appearance of edge artifacts and seams [17]. Perceptually-based approaches have also explored the assessment of subjective properties, to help improve or classify certain aspects of images. These include realism [18, 7], illumination direction [19] and aesthetic analysis of photographs [20].

While these composite improvement methods have become valuable tools, none have explicitly quantified the human response to image feature disparities - the mismatches between object and scene in composites. The manner in which the transition from realistic to unrealistic occurs for an observer, and the factors that influence it,

*e-mail: alan.dolhasz@bcu.ac.uk



Figure 2: Examples of images used in the experiments. The binary masks in the top-left corner of each image show the object chosen for processing.

are not a straightforward processes to understand. Knowledge of observer response in the context of image-based feature disparities would allow for compositing tools and matching algorithms to operate in a more optimised manner - focusing on correcting errors most perceivable to observers.

This work explores how object-scene disparities in low-level image features affect observer response to composite realism. A dataset, methodology and results are presented, showing tolerances and subjective thresholds of disparity visibility. Exposure, contrast and CCT disparities between objects and scenes are tested. Psychometric functions are then derived, to visualise response to changes in image features of the object, versus the scene.

The rest of the paper is structured as follows: Section 2 reviews approaches to quality estimation and subjective threshold quantification using perceptual data. Section 3 presents the methodology, including test design, dataset creation, participant information, experimental conditions and data analysis. Section 4 summarises the results and describes observer tolerances for each of the parameters under test. Finally, Section 5 implements these thresholds, discusses findings, reviews the impact of the dataset on the results and suggests next steps. The paper is concluded in Section 6.

2 BACKGROUND & RELATED WORK

Human judgements play an important role in application-based solutions. Aydin et al. [20] implemented subjective ratings as a method of tuning algorithms for automatic aesthetic rating of photographs. Perceptually-based models are also used in computer graphics to model visual adaptation and improve tone reproduction [21], correct properties of virtual objects in interaction tasks [22, 23], quantify image quality independent of dynamic range [24] or adapt rendering techniques to prioritise aspects most noticeable by people [25]. In order to construct such models, an understanding of the related visual processes must be developed.

To accomplish this, traditional psychophysical studies of vision often use abstract stimuli in a heavily-controlled environment, allowing for reproducibility, mathematical description, systematic variation and modelling [26]. Some argue that experimental results obtained under these controlled conditions do not generalise well to everyday visual perception in the complex, cluttered environments encountered in the real world [27, 28, 29]. Thus, methods relying on subjective quality ratings offer more flexibility, especially in the context of complex scenes, allowing to generalise results to a wider set of images. However, this comes at a cost of accuracy. Biederman, for example, [30, 31] uses an approach of violation detection in complex scenes as a method of quantifying the impact of semantic and structural scene disparities.

Subjective quality, or realism measurements involving composites and real scenes have been carried out by Xue et al. [7], to

show that offsetting image features of composites away from an ‘ideal’ state, caused decreases in subjective realism ratings, which followed a Gaussian distribution. Xue used a dataset of 20 natural images, offset both foreground and background properties and used a Likert scale. The algorithm designed by the authors performed well at improving subjective composite realism. However, the results of the perceptual study are difficult to generalise, due to the limited number of images used.

Recently, Tan et al. [32] carried out work assessing visibility of lighting inconsistencies in outdoor scenes. Their study was consistent with previous research by Lopez-Moreno et al. [13] and Ostrovsky et al. [19], specifically regarding the low sensitivity of human observers to changes in illumination direction. An exception to this were situations where the illuminant was directly behind the camera. In these cases observers found it easier to detect inconsistencies. Furthermore, the results highlight the significant impact of scene content and structure on observers’ ability to detect inconsistencies. Contrary to our work, this study addresses outdoor scenes, where the sun-sky illumination model is simpler, compared to an indoor scenario, where illuminant locations and properties (such as intensity and colour temperature) are less constrained. This makes it difficult to make direct comparisons between results.

3 MATERIALS & METHODS

The goal of the work presented here is to determine the response of human observers to feature offsets between object and scene. In order to do this for each parameter, the visibility threshold measurement method [33] and experimental conditions outlined in ITU Recommendation BT.500-13 [34] are adopted. This approach makes use of 2AFC tasks on a dataset of indoor scenes.

3.1 Feature Selection

The image features under study were *exposure*, *contrast* and *Correlated Colour Temperature (CCT)*. These features are commonly associated with images from digital cameras and they model image-based differences: exposure simulates a difference in illumination intensity, contrast models dynamic range and CCT models illuminant chromaticity differences.

Exposure changes are implemented by converting to HSV colourspace and scaling of value channel of the resulting image, as in Equation 1. The image is then converted back to sRGB for display.

$$V'(x) = aV(x, y) \quad (1)$$

Here, a is a scalar value satisfying $0.1 \leq a \leq 1.9$.

Contrast is adjusted around the middle gray level, using a point operator, threshold-based algorithm on the value channel of the HSV image [35]. The adjustment is applied to the input pixel luminance

$V(x,y)$, yielding the contrast-adjusted output pixel $V'(x,y)$, as in Equation 2.

$$V'(x,y) = b(V(x,y) - 0.5) + 0.5 \quad (2)$$

Contrast was scaled in the interval $-0.43 \leq b \leq 2.27$.

CCT: The approach of Xue *et al.* [7] is used to define the image transformations required to adjust CCT. The Robertson method [36] is used to convert from CIELUV colour space to CTY (CCT, tint and luminance). The conversion process is a table lookup and interpolation on the Planckian locus [37]. CCT is an offset using Equation 3.

$$CCT'(x,y) = CCT(x,y) + c \quad (3)$$

Where c is an offset value applied to each pixel in CCT space, whose values are restricted to $-200 \leq c \leq 200$ mired.

3.2 Dataset

A dataset consisting of 165 manually-segmented images sourced from the SUN Database [38] was created. The images were selected manually to cover a range of indoor scenes and objects (see Figure 2 for examples). The horizontal resolution of all images was normalised to 500px, in order to fit three instances of each image on a 1920x1080px display. By using natural images, segmenting objects within them, and systematically altering image properties of those objects, composites with controlled disparities can be simulated. This approach leaves original position [31], semantics [30], illumination [12] and reflections [2] of the objects unchanged.

The processing applied consisted of either an offset (CCT) or scaling (exposure, contrast) of one of the image features in the segmented area of the image (representing an object). The severity of the offset was selected from a range of 11 values, ranging from negative (e.g. reduction of contrast) through null (no processing, identical to original image) to positive (e.g. increase of contrast). Equal distribution of stimulus levels across images in the dataset was also ensured to minimise the bias induced by variation of image content. Based on [7, 39], an unaltered photograph is an example of an ideal composite, whereby the statistical properties of all image features across the object-scene combination are in alignment.

3.3 Apparatus and Task

A self-calibrating Eizo ColorEdge CG247 monitor was used, calibrated to an sRGB profile under the experimental conditions (see Fig. 3). Observers were required to view 165 image triples in a 2AFC setting. Each triple consisted of an original, unprocessed image, a binary image indicating the location of the object, and a processed version of this image (see Fig. 3). Including object location mitigates any lapses due to observers searching for the object, or completing the task based on the wrong object. In alignment with [34], the images were displayed on an sRGB middle gray background and observers were given 10 seconds to view each image pair, followed by 5 seconds to cast their vote. The task was to indicate which of the two colour images looked more realistic.

3.4 Experimental Design

A total of 75 observers, 33 female, mean age of 28.53 (SD=10.54), were recruited from a population of university staff and students. All observers were volunteers and were not rewarded in any way. Observers were then evenly and randomly distributed into three groups, one for each of the three experiments. The selection process ensured that each user was presented with each of the 11 stimulus levels (see Fig.1) an equal number of times. Experiment I used exposure, II used contrast and III used CCT. Each experiment lasted 40 minutes on average, resulting in around 50 hours of testing.

Each observer was tested for normal or corrected-to-normal visual acuity and colour vision using a Snellen chart and Ishihara test,

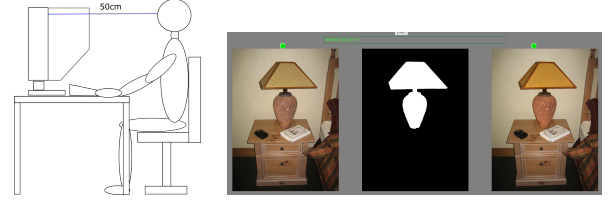


Figure 3: An illustration of the experimental setup (left) and an example of the test screen as seen in the experiments (right).

respectively. Observers with a visual acuity below 0.8, or those suffering from colour blindness were rejected. Observers were positioned 50cm away from the display and asked to familiarise themselves with the instructions, informed that in each presentation, the task would consist of selecting the most realistic image from the two stimuli and that the stimulus range would vary throughout the experiment. Finally, observers were given an opportunity to ask the experimenter questions. During the experiment, observers were given the choice between using the keyboard or the mouse to vote.

3.5 Analysis of Results

The analysis of experimental results follows the recommendations of ITU Report BT. 1082-1 [33] and the procedures detailed by Wichmann & Hill (2001) [40, 41]. First, proportions of correct responses per stimulus value are calculated. Here, “correct” is defined as selecting the original image, as opposed to the processed image. Psychometric functions (PFs) are fit to the resulting data points using the *Psignifit* toolbox version 3.0 for Matlab [42], which implements the maximum-likelihood method presented in [40]. The psychometric function $\psi(x)$ describes the relationship between the probability of a correct response p , and a given stimulus intensity x . This is commonly denoted as in Equation 4:

$$\psi(x; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)F(x; \alpha, \beta) \quad (4)$$

In this study, $F(x; \alpha, \beta)$ is a logistic function, as in Equation 5:

$$F(x; \alpha, \beta) = \frac{1}{1 + \exp(-\frac{x - \alpha}{\beta})} \quad (5)$$

The parameters $\alpha, \beta, \gamma, \lambda$ of ψ define the shape of the curve, and correspond respectively to its *threshold*, *slope*, lower (*guess rate*) and upper (*lapse rate*) asymptote. The threshold α of this psychometric function describes its displacement along the abscissa. Specifically, it marks the stimulus intensity, for which the probability of a correct response is the same, as that of a guess. Assuming that $\gamma = 0.5$ and $\lambda = 0$, α corresponds to the stimulus value yielding a .75 proportion of correct responses. The slope β describes the steepness, or gradient of the function, at the threshold α . In the 2AFC scenario, the guess rate γ is fixed to 0.5, as the probability of a correct guess in an n-alternative setting is $1/n$. λ represents the probability of a stimulus-independent lapse - an incorrect response, despite an arbitrarily high stimulus intensity.

The fitting process is carried out using the *Pool-then-fit* method, adopted from Wallis et al. (2013) [43]. Next, 95% confidence intervals (CIs) for all parameters are found using the bias-corrected and accelerated (BC_a) bootstrap method [44], as suggested by Hill (2011) [45].

4 RESULTS

This section details the results for the three experiments. A total of $165 \times 25 \times 3 = 12375$ data points were gathered. To ensure that each image-offset combination was presented at least twice, 25 observers were shown 165 images, with each of the 11 offsets being presented 15 times ($25 \times 165 = 4125$). Here, we report on

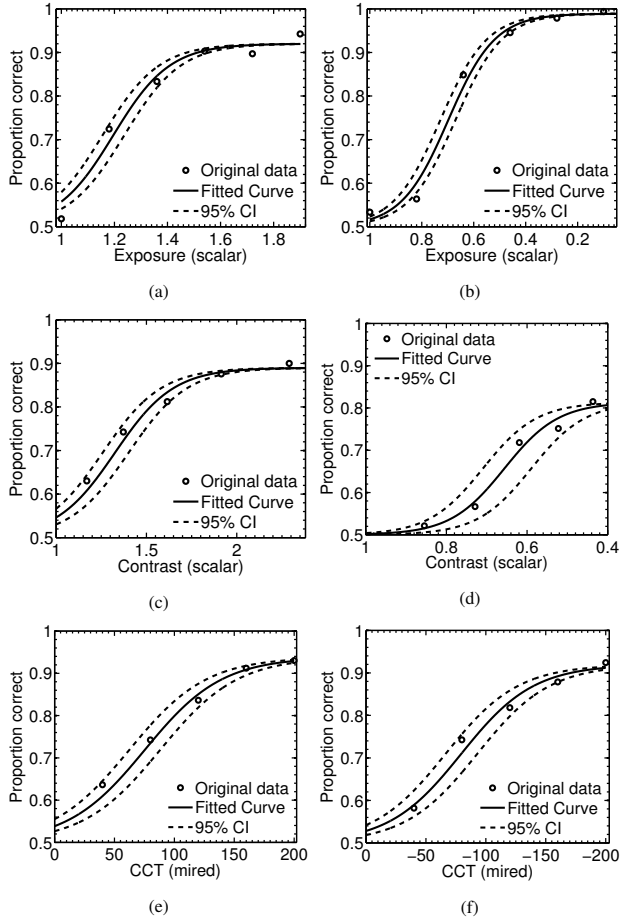


Figure 4: Psychometric functions fitted to the data obtained for positive (left) and negative (right) feature offsets of exposure (a & b), contrast (c & d) and CCT (e & f). The solid curves represent the fitted psychometric functions, dashed curves represent the 95% CIs for threshold. Circles represent the original data. Parameters for these psychometric functions can be seen in Table 1.

$165 \times 11 \times 2 = 3630$ data points from those collected for each experiment, thus ensuring two responses for each image-offset combination are included. Psychometric functions fitted to responses for positive and negative feature offsets from Experiments I, II and III can be seen in Figure 4. They reflect the measures of thresholds, slopes and lapse rates obtained using the maximum-likelihood procedure and pooling method described in Section 3.5. Individual threshold, slope and lapse rate measures for each experiment, as well as the corresponding 95% confidence intervals can be found in Table 1.

Exposure: The thresholds for positive and negative exposure offsets were 1.1953, 95% CI [1.1557, 1.2365] and 0.6852, 95% CI [0.6572, 0.7154] respectively. Slopes for the threshold points obtained were 0.1049, 95% CI [0.0695, 0.1495] for positive exposure offsets and 0.0885 [0.0664, 0.1141] for negative ones. As Figures 4a and 4b and Table 1 show, the slope for negative offsets is steeper, than that for positive offsets, indicating a narrower exposure offset interval, in which the transition from realistic to unrealistic composite rating occurs. Lapse rates were smaller for negative exposure offsets, at 0.0093, 95% CI [0.0021, 0.0183], compared to positive exposure offsets, at 0.0640, 95% CI [0.0440, 0.0834]. The lapse rate estimates are larger for positive offsets, as seen in Table 1.

Exposure	Positive Offset		Negative Offset	
	Value	95% CI	Value	95% CI
Threshold	1.1953	[1.1557, 1.2365]	0.6852	[0.6572, 0.7154]
Slope	0.1049	[0.0695, 0.1495]	0.0885	[0.0664, 0.1141]
Lapse Rate	0.0789	[0.0570, 0.0978]	0.0102	[0.0000, 0.0196]
Contrast	Positive Offset		Negative Offset	
	Value	95% CI	Value	95% CI
Threshold	1.3249	[1.2590, 1.3905]	0.6678	[0.6433, 0.7527]
Slope	0.1451	[0.0955, 0.2070]	0.0426	[0.0250, 0.1601]
Lapse Rate	0.1139	[0.0852, 0.1394]	0.2070	[0.1532, 0.2404]
CCT	Positive Offset		Negative Offset	
	Value	95% CI	Value	95% CI
Threshold	72.172	[60.071, 84.414]	84.339	[71.011, 99.538]
Slope	29.412	[20.063, 41.146]	32.053	[22.057, 46.216]
Lapse Rate	0.0756	[0.0360, 0.1029]	0.0756	[0.0215, 0.1099]

Table 1: Results, detailing threshold, slope and lapse rate estimates for Experiments I (exposure), II (contrast) and III (CCT)

Contrast: The thresholds obtained were 1.3249, 95% CI [1.2590, 1.3905] for positive contrast offsets and 0.6678, 95% CI [0.6433, 0.7527] for negative ones. It was found that observers' responses for negative contrast offsets varied the most out of the experiments. The lapse rate for those offsets was 0.2070, 95% CI [0.1532, 0.2404], while for positive offsets this was 0.1139, 95% CI [0.0852, 0.1394]. Slope values were 0.1451, 95% CI [0.0955, 0.2070] for positive offsets, and 0.0426, 95% CI [0.0250, 0.1601] for negative ones.

CCT: For CCT, thresholds obtained from these results are 72.172, 95% CI [60.071, 84.414] mired for positive CCT offsets and 84.339, 95% CI [71.011, 99.538] mired for negative offsets. Slope values for these thresholds are 29.412, 95% CI [20.063, 41.146] for positive CCT offsets and 32.053, 95% CI [22.057, 46.216] for negative CCT offsets. Lapse rates across the positive and negative offset conditions were similar, at 0.0756, 95% CI [0.0360, 0.1029] and 0.0756, 95% CI [0.0215, 0.1099] respectively.

5 DISCUSSION

Overall, observer response followed a similar trend, as described by Xue [7] - realism ratings decrease as the object-scene disparities increase. While it is not possible to compare performance between features, several noteworthy trends were found in the experimental data obtained. Contrast offsets, particularly negative ones, yielded highest lapse rate estimates and widest CIs (see Fig. 4c and 4d). A possible explanation for this outcome can be sought in the work of Marius 't Hart et al. [46], who shows that contrast is correlated with attention, and decreases in contrast, particularly those applied to object regions, reduce fixation and detection probability. The same is not true for relative increases of contrast, which is reinforced in the work presented here. Thus, when matching contrast of an object to that of a scene, underestimating object contrast is likely to appear less realistic to an observer than overestimating it, which will in turn attract attention to that object, as found in [46].

Both CCT and Exposure covered the ranges adequately, receiving 100% correct responses for the highest offsets, in the case of some observers. As CCT was offset in the perceptually-uniform space of mired, the resulting PFs are expectedly similar. Additionally, it seems that some extreme CCT offsets can be interpreted as plausible differences in object reflectance, increasing the lapse rates, while still appearing realistic. The variability in responses for each offset level can be attributed to image and object changes. This is consistent with studies by Tan [32] and Xue [7], who found significant differences between the consistency of ratings for different images, as well as across participants. Through the use of a larger dataset of 165 images, our work also indicates how much variability can be expected across general composites. This subjectivity of realism judgements is further illustrated by the confidence intervals

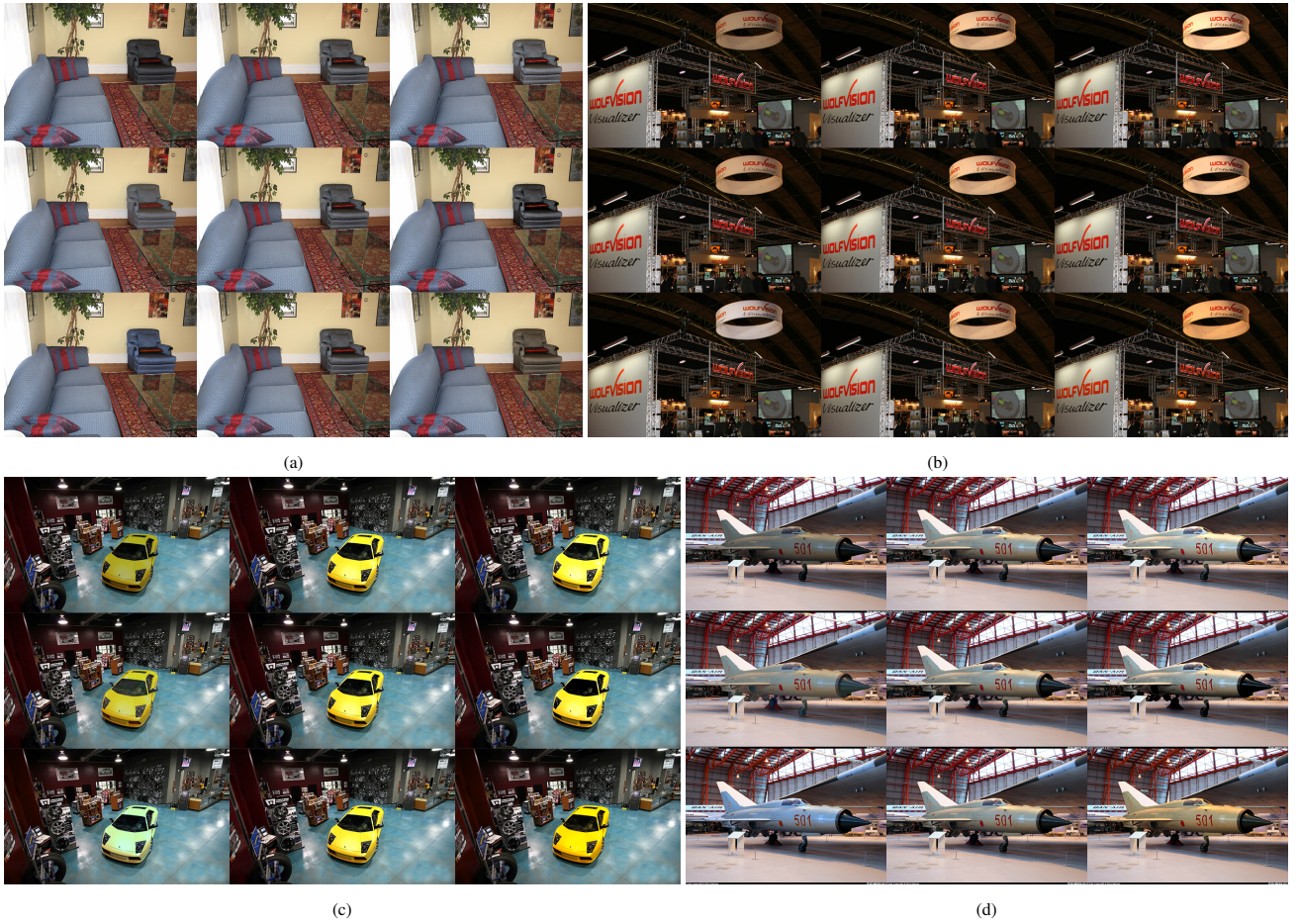


Figure 5: Applying the resulting visibility thresholds for negative (left column) and positive (right column) offsets of exposure (top row), contrast (middle row) and CCT (bottom row). The original images are shown in middle columns.

in Figure 4.

The results of the experiments were also applied to a set of images from the original dataset, for purpose of visualisation. Objects within these images were processed using the resulting threshold values from Table 1 (in bold). These images can be seen in Figure 5 (for optimal quality, please view the digital version of the paper). Here, each 3x3 image grid compares the effects of applying the negative (left column) and positive offset thresholds (right column) to the original image (middle column). The rows from top to bottom represent exposure, contrast and CCT, respectively.

The threshold images are visibly different from the originals, indicating that realism responses were not a function of sheer impairment visibility. Rather, they are a result of a more complex process, based on inference and influenced by preference and properties of complex scenes [28, 29]. This is also supported by the results of [32], who found significant differences in observer responses for the same disparities, across different scenes and objects. In some cases, even in an ideal scenario where both the reference and a modified version of an image are present, and the location of the modified object is specified, observers often select the real image as the ‘unrealistic’ one. The interesting question arising here is: “How are these preferential responses influenced by the scene content?”. We aim to address this in our future work. Interestingly, several observers reported difficulty in predicting the exact hue of monochromatic objects, when CCT was offset. Another comment

included difficulty in judging correct exposure when objects were nearby strong lights. These comments can be investigated formally as part of further studies, which include comparison of psychometric functions between controlled and uncontrolled environments, extension of the study to include eye-tracking, as a measure of visual task difficulty, as well as implementing this experiment into AR video see-through devices in order to ascertain changes to realism thresholds.

6 CONCLUSIONS

This paper has presented the results of three experiments analysing responses to image-based object-scene disparities. Resulting PFs indicate generalised correspondences between feature offsets and observer realism responses. A degree of subjectivity in terms of realism ratings across observers and scenes was also observed. While a detailed comparison with existing work is not possible due to methodological differences, the results are consistent with previous work in this area and provide a starting point for further research into perception of realism. Thresholds indicating the feature offsets at which observers transition from a ‘realistic’ to ‘unrealistic’ response and their confidence intervals have also been provided, they are 0.6852 and 1.1953 for exposure, 0.6678 and 1.3249 for contrast and -84.339 and 72.172 for CCT. These results can be taken forward to help create realistic composites. Furthermore, while carried out in a controlled lab setting, the methodology and results of this study are applicable to any AR/MR

scenario where the augmentation and scene share a single screen. The image dataset and corresponding responses for each experiment described in this paper is available at <http://dmtlab.bcu.ac.uk/composites>.

REFERENCES

- [1] S. Wright, *Digital compositing for film and video*. Taylor & Francis, 2013.
- [2] P. Debevec, "Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography," in *ACM SIGGRAPH 2008 classes*. ACM, 2008, p. 32.
- [3] S. Prince, *Digital visual effects in cinema: the seduction of reality*. Rutgers University Press, 2011.
- [4] W. Barfield, *Fundamentals of wearable computers and augmented reality*. CRC Press, 2015.
- [5] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre, "Recent advances in augmented reality," *Computer Graphics and Applications, IEEE*, vol. 21, no. 6, pp. 34–47, 2001.
- [6] J.-Y. Zhu, P. Krahenbuhl, E. Shechtman, and A. A. Efros, "Learning a discriminative model for the perception of realism in composite images," in *Proceedings of the IEEE ICCV*, 2015, pp. 3943–3951.
- [7] S. Xue, A. Agarwala, J. Dorsey, and H. Rushmeier, "Understanding and improving the realism of image composites," *ACM TOG*, vol. 31, no. 4, p. 84, 2012.
- [8] E. Kruijff, J. E. Swan II, and S. Feiner, "Perceptual issues in augmented reality revisited," in *ISMAR*, vol. 9, 2010, pp. 3–12.
- [9] T. Pouli and E. Reinhard, "Progressive color transfer for images of arbitrary dynamic range," *Computers & Graphics*, vol. 35, no. 1, pp. 67–80, 2011.
- [10] E. Reinhard, A. O. Akyuz, M. Colbert, C. E. Hughes, and M. O'Connor, "Real-time color blending of rendered and captured video."
- [11] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, no. 5, pp. 34–41, 2001.
- [12] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sittig, and D. Forsyth, "Automatic scene inference for 3d object compositing," *ACM TOG*, vol. 33, no. 3, p. 32, 2014.
- [13] J. Lopez-Moreno, S. Hadap, E. Reinhard, and D. Gutierrez, "Compositing images through light source detection," *Computers & Graphics*, vol. 34, no. 6, pp. 698–707, 2010.
- [14] J.-F. Lalonde, A. Efros, S. G. Narasimhan *et al.*, "Estimating natural illumination from a single outdoor image," in *ICCV, 2009 IEEE 12th*. IEEE, 2009, pp. 183–190.
- [15] K. Rohmer, W. Buschel, R. Dachsel, and T. Grosch, "Interactive near-field illumination for photorealistic augmented reality on mobile devices," in *2014 IEEE ISMAR*, Sept 2014, pp. 29–38.
- [16] K. Sunkavalli, M. K. Johnson, W. Matusik, and H. Pfister, "Multi-scale image harmonization," in *ACM TOG*, vol. 29, no. 4. ACM, 2010, p. 125.
- [17] J. Jia, J. Sun, C.-K. Tang, and H.-Y. Shum, "Drag-and-drop pasting," *ACM TOG*, vol. 25, no. 3, pp. 631–637, 2006.
- [18] J.-F. Lalonde, A. Efros *et al.*, "Using color compatibility for assessing image realism," in *ICCV 2007, IEEE 11th*. IEEE, 2007, pp. 1–8.
- [19] Y. Ostrovsky, P. Cavanagh, and P. Sinha, "Perceiving illumination inconsistencies in scenes," 2001.
- [20] T. O. Aydin, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *IEEE TVCG*, vol. 21, no. 1, pp. 31–42, 2015.
- [21] J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg, "A model of visual adaptation for realistic image synthesis," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 249–258.
- [22] G. Hough, I. Williams, and C. Athwal, "Fidelity and plausibility of bimanual interaction in mixed reality," *IEEE TVCG*, vol. 21, no. 12, pp. 1377–1389, 2015.
- [23] —, "Measurements of live actor motion in mixed reality interaction," in *2014 IEEE ISMAR*. IEEE, 2014, pp. 99–104.
- [24] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Dynamic range independent image quality assessment," in *ACM TOG*, vol. 27, no. 3. ACM, 2008, p. 69.
- [25] C. O'Sullivan, S. Howlett, Y. Morvan, R. McDonnell, and K. O'Connor, "Perceptually adaptive graphics," *Eurographics state of the art reports*, vol. 4, pp. 1–24, 2004.
- [26] D. Bartz, D. Cunningham, J. Fischer, and C. Wallraven, "The role of perception for computer graphics," *Eurographics state-of-the-art reports*, pp. 65–86, 2008.
- [27] J. J. Gibson, *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.
- [28] A. Yuille and D. Kersten, "Vision as bayesian inference: analysis by synthesis?" *Trends in cognitive sciences*, vol. 10, no. 7, pp. 301–308, 2006.
- [29] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as bayesian inference," *Annu. Rev. Psychol.*, vol. 55, pp. 271–304, 2004.
- [30] I. Biederman, *On the semantics of a glance at a scene*, 1981.
- [31] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cognitive psychology*, vol. 14, no. 2, pp. 143–177, 1982.
- [32] M. Tan, J.-F. Lalonde, L. Sharan, H. Rushmeier, and C. O'Sullivan, "The perception of lighting inconsistencies in composite outdoor scenes," *ACM Transactions on Applied Perception (TAP)*, vol. 12, no. 4, p. 18, 2015.
- [33] C. X. P. Assembly, "Report 1082-1, studies toward the unification of picture assessment methodology," *Reports of the CCIR, XI-Part*, vol. 1, pp. 384–414, 1990.
- [34] I. Rec, "Bt. 500-13," *Methodology for the subjective assessment of the quality of television pictures*, 2012.
- [35] J. C. Russ, *The image processing handbook*. CRC press, 2011.
- [36] A. R. Robertson, "Computation of correlated color temperature and distribution temperature," *JOSA*, vol. 58, no. 11, pp. 1528–1535, 1968.
- [37] G. Wyszecki and W. S. Stiles, *Color science*. Wiley New York, 1982, vol. 8.
- [38] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR, 2010 IEEE conference on*. IEEE, 2010, pp. 3485–3492.
- [39] S. Prince, "True lies," *Film Theory: Critical Concepts in Media and Cultural Studies*, vol. 4, no. 3, p. 85, 2004.
- [40] F. A. Wichmann and N. J. Hill, "The psychometric function: I. fitting, sampling, and goodness of fit," *Perception & psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.
- [41] —, "The psychometric function: II. bootstrap-based confidence intervals and sampling," *Perception & psychophysics*, vol. 63, no. 8, pp. 1314–1329, 2001.
- [42] I. Fründ, N. V. Hanel, and F. A. Wichmann, "Inference for psychometric functions in the presence of nonstationary behavior," *Journal of Vision*, vol. 11, no. 6, May 2011. [Online]. Available: <http://www.journalofvision.org/content/11/6/16>
- [43] S. A. Wallis, D. H. Baker, T. S. Meese, and M. A. Georgeson, "The slope of the psychometric function and non-stationarity of thresholds in spatiotemporal contrast vision," *Vision research*, vol. 76, pp. 1–10, 2013.
- [44] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical science*, pp. 54–75, 1986.
- [45] N. J. Hill, "Testing hypotheses about psychometric functions," Ph.D. dissertation, University of Oxford, 2001.
- [46] B. Marius't Hart, H. C. E. F. Schmidt, I. Klein-Harmeyer, and W. Einhäuser, "Attention in natural scenes: contrast affects rapid visual processing and fixations alike," *Phil. Trans. R. Soc. B*, vol. 368, no. 1628, p. 20130067, 2013.

[POSTER] Composite Realism: Effects of Object Knowledge and Mismatched Feature Type on Observer Gaze and Subjective Quality

Alan Dolhasz*

Maite Frutos-Pascual†

Ian Williams‡

DMT Lab
Birmingham City University

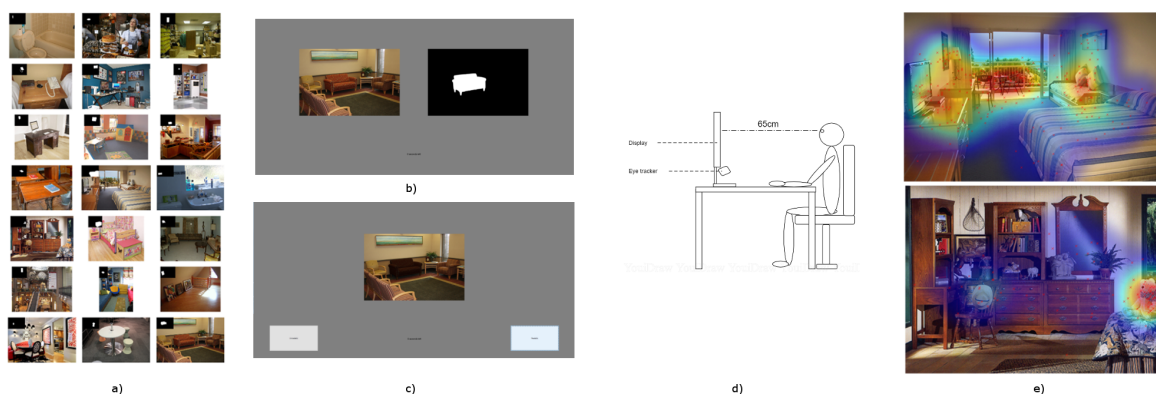


Figure 1: **a)** Images and segmentation masks used in our experiment (selected from [3]). Binary masks indicating object location are overlaid in the top left corner (scaled down for display purposes); **b)** Example of reference image with binary mask indicating object location; **c)** Example of test image with processed object; **d)** Setup of experimental apparatus. Observers positioned at a distance of 65cm from height-adjusted display. Eye tracker facing towards observer's eyes; **e)** Examples of heatmaps visualising average fixation maps across observers for a given image.

ABSTRACT

We report on the results of the first visual search and rating study (N=60) evaluating human gaze when assessing the realism of image composites. The effects of object identity knowledge and mismatched feature type on observers' gaze and subjective realism scores are studied. Gaze metrics used include: fixation count, fixation duration, time and duration of first fixation on target object, as well as area of interest similarity and inter-observer consistency. Monte-Carlo-based techniques are used for analysis of the data obtained. Results indicate that while knowledge of object identity impacts gaze allocation and response times, it leaves subjective realism ratings unchanged. We show that the type of mismatched feature (correlated colour temperature vs exposure) has a significant impact on fixation counts and durations. This study provides a first step to utilising objective gaze metrics to better understand subjective assessment processes and leads towards the development of gaze-inspired compositing methods.

Keywords: compositing, eye tracking, subjective quality, realism

Index Terms: Human-centred Computing [Visualisation]: Empirical studies in visualization; Computing methodologies [Computer Graphics]: Graphics systems and interfaces—Perception

1 INTRODUCTION

Compositing is a process aiming to combine multiple images or parts thereof into a seamless whole. Aside from fundamentally un-

derpinning mixed and augmented reality (AR & MR), this process is commonplace in a range of application domains, such as film or visual effects. Depending on the application, compositing is either carried out automatically, in real-time (e.g. when placing virtual objects over a camera image of the real world in AR) or manually, in an offline process (e.g. when compositing visual effects for film).

The aim of compositing is to create a believable, realistic result, through minimising the noticeable disparities between the individual elements of the composite. However, in this context the resulting realism is usually subjective and affected by the type and severity of mismatches (e.g. illumination, orientation, semantics etc.), as well as the final application context. As opposed to more traditional uses of compositing, in AR and MR observers/users are usually explicitly aware of the identity of the composited elements, due to application and implementation differences.

A recent study [3] showed that subjective realism ratings for composite feature mismatches tend to be inversely correlated with their severity. Furthermore, scene context was shown to have a large impact on some of the realism ratings. However, this work [3] does not explain whether the prior knowledge of the identity of the mismatched object had an effect on subjective realism scores or how the scene impacts the realism judgement.

Here, we present the results of the first gaze-based study assessing the impact of object identity knowledge and type of mismatched feature on observer gaze and subjective realism ratings. The rest of the paper is structured as follows: Section 2 discusses relevant background concepts and related work. Section 3 describes our methodology, experimental setup and analytical methods. Section 4 summarises results and findings. Finally, Section 5 summarises our conclusions and suggests future work directions.

*e-mail:alan.dolhasz@bcu.ac.uk

†e-mail:maite.frutos@bcu.ac.uk

‡e-mail:ian.williams@bcu.ac.uk

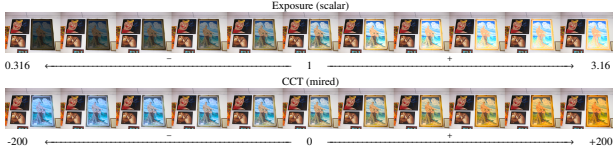


Figure 2: Offsets applied to segmented objects in test images. Top row: exposure offsets (scalar multiplication); Bottom row: CCT offsets in mired (subtraction / addition)

2 RELATED WORK

2.1 Perception of Disparities

Human visual perception is driven by visual attention (VA). There is evidence that perceptual thresholds can be affected by attention [17]. Distortions or disparities in salient regions are more likely to contribute to a lower subjective quality score than those in non-salient regions [7]. Furthermore, in MR/AR applications, the identity of the augmented content is often known and thus VA is focused on it. Subjective quality, or realism, of composite scenes is influenced by noticeable mismatches between features of the inserted object and those of the rest of the scene [3]. While some blatantly disparate features can go unnoticed [2], other mismatches are noticeable with minimal conscious effort [1]. Also, different types of distortion impact subjective quality scores differently [22, 31].

The deployment of VA is facilitated by two distinct mechanisms. Bottom-up VA relies on preattentive vision [29]. It is transient, involuntary, largely driven by saliency and independent of task. Top-down VA is deployed in a voluntary manner, is task-dependent and driven by higher level factors, such as semantics, context, preference, expectations, emotions and experience [32].

2.2 Eye Movement Metrics & Subjective Ratings

Human eye movements can provide objective information complementary to conventional subjective ratings, such as questionnaires or rating scales [6]. As eye movements are paramount to acquisition of visual information while performing cognitive tasks, studying how they are deployed can reveal visual strategy and features relied upon during the completion of a task [4].

Examples of the use of gaze data in the assessment of visual strategy and attention exist in both free-viewing conditions [32] and specific tasks such as reading [23], visual search [19], objective image quality metrics [14], decision-making [17], scene perception [10] and subjective quality evaluation [27]. However, there are few cases of gaze data used in subjective evaluation of visual realism.

Zangemeister et al. [33] used eye tracking to analyse visual strategies when viewing abstract and realistic art. Ninassi et al. [15] used objective eye metrics to study the impact of task on VA in subjective image quality assessment through comparing empirical fixation maps. Elhelw et al. [6], studied the impact of different image features on the perceived realism of real and synthetic bronchoscopes images. Finally, Vu et al. [28] assessed the impact of common global image distortions, such as blurring, noise, packet loss and JPEG compression artifacts on fixation patterns.

Gaze-based approaches utilising objective metrics have not yet been employed to study perception of object-scene mismatches in the context of composite realism. Particularly, in a manner which would allow generalisability or transferability of the findings. In this paper, we wish to address these limitations through a study of human gaze when assessing the quality and realism of composites in a first step towards developing observer models for automated compositing tools.

Table 1: Experiment design, showing 2×2 factorial design and assigned observer groups.

		Object location & Identity	
		No location	Location
Mismatched Feature	Exposure	Group A (EN)	Group B (EL)
	CCT	Group C (CN)	Group D (CL)

3 METHOD

3.1 Overview & Design

We apply a between-groups 2×2 factorial design. The factors were:

- *mismatched object feature*: exposure (E) and correlated colour temperature (C)
- *observer knowledge of the location/identity of the processed object*: no location (N) and location (L)

Using these factors, we define 4 experimental conditions (see Table 1): **EN** (exposure, no location), **EL** (exposure, location), **CN** (CCT, no location), **CL** (CCT, location). The experimental procedure adopted is an adaptation of the *double-stimulus impairment scale* (DSIS) method [21]. In each trial, observers had to assess if an object-scene combination appeared realistic or unrealistic.

3.2 Stimuli

The experiments used 33 images with segmented objects, selected from a subset of the SUN Dataset, as in [3]. The images were selected to cover a range of object types and luminance values, according to the mean luminance of the segmented objects, (in CIE $L^*a^*b^*$ colour space). The selection was made so that the segmented object occupied no more than 1/3 of the total image area. The horizontal resolution of the images was normalised to 600 pixels (px), preserving the aspect ratio. At a 65 cm viewing distance 37 px on the screen corresponded to 1° visual angle (VAn). Examples of these images can be seen in Figure 1a.

In each image, the segmented object had either an exposure or CCT offset applied to it to simulate an unrealistic combination of object and scene ([31, 3]). Exposure offsets were implemented using a scaling of the V channel in HSV space, whereas CCT offsets were implemented using an additive offset in the perceptually uniform *mired* space, following the method described by Robertson [25]. To minimise learning effects, the offsets applied to the objects were varied: exposure was scaled in 11 logarithmically spaced steps between .3162 and 3.162, corresponding to a range of -1.661 to 1.661 in \log_2 domain. CCT was offset in 11 increments of 40 mired, between -200 and 200 mired. The offset ranges followed those in: [31, 3]. The order of relative offset intensities was kept the same for exposure and CCT. Examples of offsets applied to an image can be seen in Figure 2.

3.3 Observers

Sixty (60) observers, recruited from a population of university staff and students were randomly assigned into 4 groups (see Table 1). All observers were volunteers and were not rewarded. The following groups were compiled: *Group A* (condition EN) with a mean age of 26.00 ($SD = 5.76$) 7 females, *Group B* (condition EL) with a mean age of 25.93 ($SD = 4.32$), 7 females, *Group C* (condition CN) with a mean age of 28.47 ($SD = 4.81$) 6 females and *Group D* (condition CL) with a mean age of 32.27 ($SD = 8.15$) 7 females. All observers had normal or corrected-to-normal vision and normal colour vision, as verified by a SNELLEN chart and Ishihara test. Each observer gave consent to take part in the experiment and was naïve to its purpose.

3.4 Apparatus

3.4.1 Display

Images were displayed on a 22" 60 Hz Iiyama ProLite B2280HS LED monitor, calibrated to sRGB colour space using an X-Rite i1 Display Pro calibrator. The monitor was placed in a evenly illuminated room and the calibration was corrected for both the chromaticity and intensity of the ambient illumination. The maximum measured luminance level of the display was 214 cd/m^2 , while the black luminance was $.375 \text{ cd/m}^2$. When displayed, the images occupied $11.8^\circ \times 7.9^\circ$ VAn.

3.4.2 Eye tracker

We used a Tobii X1 Light eye tracker, fixed below the display at a distance of 65 cm from the observers' head (see Figure d), as recommended by the manufacturer. Average binocular accuracy, as reported by the manufacturer, was $.4^\circ$ VAn and an average precision was $.2^\circ$ VAn at the selected viewing distance. Its typical sampling rates fall between 28-32 Hz. The eye tracker compensated for head movements of up to 44 cm horizontally and 32 cm vertically, which removed the need for a chin rest. The device was recalibrated for each observer, following the manufacturer's recommendations [26].

3.5 Procedure

3.5.1 Preparation

Observers were asked to familiarise themselves with the test instructions and shown examples of the reference (original unmodified - see Figure 1b) and test images (processed object - see Figure 1c). Observers were then given an opportunity to ask questions.

3.5.2 Trials

Three predefined sets of 11 trials were administered per observer, 11 images per set, with rest breaks between. During each trial observers in conditions *EN* (Group A) and *CN* (Group C) would first see a reference image, which they were instructed to analyse. In the case of two conditions which revealed the location / identity of the processed object (*EL* - Group B and *CL* - Group D), a binary mask was also displayed next to the reference image (see Figure 1b). This was displayed for 10 seconds, followed by a 3-second middle grey screen to ensure change blindness [30]. Next, the test image was displayed (see Figure 1c). Here, a feature offset had been applied to the segmented object and the observers' task was to decide whether this test image appeared realistic or unrealistic, compared to the reference image. Observers had 10 seconds to analyse the image and click the button corresponding to their chosen answer. They were also requested to respond as quickly and accurately as possible. This procedure was repeated for each observer.

3.6 Analysis

3.6.1 Fixation Extraction and Fixation Maps

Fixations and saccades were extracted using the ClusterFix package for MATLAB [12]. Resulting fixation locations were then mapped onto fixation maps (sparse 2D histograms) \mathbf{F}_{obs} , for each image and observer. Only fixation data from the image region is used in this process. Fixations falling outside of the reference and test image regions (such as the mask image or buttons) were rejected. Joint fixation maps \mathbf{F}_{joint} were generated by normalising and averaging fixations for each image across all observers from a single condition:

$$\mathbf{F}_{joint} = \frac{1}{N} \sum_{o=1}^N \frac{\mathbf{F}_{obs}}{\sum_{i=1}^n \sum_{j=1}^m \mathbf{F}_{ij}^{obs}} \quad (1)$$

Here, \mathbf{F}_{ij}^{obs} denotes element at row i and column j of fixation map \mathbf{F}_{obs} . We normalise each joint histogram by the sum of its elements in order to avoid biasing the joint fixation maps towards observers

who executed a higher fixation count. Thus, each bin of the fixation map represents a proportion of task time that location was fixated by that observer.

3.6.2 Eye Movement Metrics

We use commonly applied eye movement metrics, as in [8]: fixation count (Fc), fixation duration (Fd), time to first fixation on processed object ($TFFO$) and duration of first fixation on processed object ($DFFO$). Fc correlate positively with the amount of information to be attended and task difficulty. Fd relate to the usefulness of particular regions to task completion and overall difficulty of information extraction [9]. A distinction must be made between early fixations, driven by bottom-up VA mechanisms and later fixations, driven by top-down mechanisms. Shorter $TFFO$ can be an indicator of an object attracting bottom up VA [11] and longer $DFFO$ can point to an object's task-relevant semantic informativeness [9]. Fd can be affected by scene context: objects that do not belong in the scene tend to attract longer fixations than object that do [20].

3.6.3 Fixation Map Metrics

To assess similarity between observers' fixation distributions within one condition we use Inter-Observer Consistency (IOC) [13], specifically the "one against all" approach. This compares the fixation map of each observer against a joint fixation map of all other observers using a similarity metric. We also adopt the Area of Interest Similarity (*AOIS*), which expresses the degree of similarity between joint fixation maps across our experimental conditions.

To calculate the similarity between joint fixation maps, for both *IOC* and *AOIS*, we use the similarity score (SS) as recommended by Riche et al. [24]. This approach computes the sum of the minima between each point of two probability distributions:

$$SS(P, R) = \sum_{i,j} \min(P_{i,j}, R_{i,j}) \quad \text{where} \quad \sum_{i,j} P_{i,j} = \sum_{i,j} R_{i,j} = 1.0 \quad (2)$$

P and R represent discrete 2D probability distributions (PDs). We convert discrete fixation maps to PDs by placing a Gaussian with $\sigma = 1^\circ$ VAn at the location of each fixation in order to model uncertainty in viewing location caused by the accuracy and precision of the eye tracker, as in [13]. See Figure 1e for a heat map visualisation of these fixation distributions, along with the original fixations, marked as red 'x' symbols.

3.6.4 Statistical Measures

To estimate parameters and compare between conditions we use the nonparametric bootstrap method proposed by [5]. Bootstrapping is used both to estimate the means/medians of the eye movement metrics from the empirical samples, as well as calculate their 95% confidence intervals (CIs), standard errors and bias. We use the bias corrected and accelerated (BCa) method to calculate CIs [18]. Furthermore we use bootstrapping to compare assess group differences and effect sizes. Also, we adopt Fisher's permutation test [16] as a means for testing the statistical significance of differences between groups. The chosen statistic for this procedure is the difference of means ($\bar{x}_1 - \bar{x}_2$), unless the empirical data distribution is heavily skewed - in such cases we use the difference of medians. The number of simulated samples for our bootstrap procedures and permutation tests is 5000. When computing correlations across realism responses, we use Pearson correlation.

3.7 Hypotheses

The goal of this study is to identify whether the mismatched feature of a composite as well as observers' knowledge of the mismatched object identity/location influence their realism ratings and gaze allocation. Furthermore we wish to understand whether interactions exist between the two evaluated factors. We test the null hypotheses

that varying the mismatched object feature and knowledge of object location/identity will have no effect on:

Eye Movement Metrics:

- H1: Fixation counts (F_c)
- H2: Fixation durations (F_d)
- H3: Time to first fixation on target object ($TFFO$)
- H4: Duration of first fixation on object ($DFFO$)

Fixation Map Metrics:

- H5: Inter-Observer Consistency (IOC)
- H6: Area of Interest Similarity ($AOIS$)

Realism Responses:

In addition to the gaze-based hypotheses, we make the following hypotheses for observers' realism ratings and response times:

- H7: RTs will be shorter when object location is known.
- H8: Realism ratings will not change across location conditions.

4 RESULTS & FINDINGS

We present the results of our experiments below, detailing the implication for each individual metric. Figure 3 shows the bootstrapped mean/median values for all metrics under study, along with their respective 95% CIs. Figure 4 shows bootstrapped pairwise differences between the means/medians of each evaluated metric. Figure 5 shows AOIS values between conditions.

4.1 Eye Movement Metrics

4.1.1 Fixation Count (F_c)

Results: Fisher's permutation test indicated that mean F_c values (see Fig.3) obtained for the *CN* and *EN* conditions were significantly different from every other condition at $p < .05$ (see Fig.4). In contrast, when object location was revealed in conditions *EL* and *CL*, significant differences in F_c were not found ($p = 0.81$). Largest effect sizes were observed when comparing mean F_c values between conditions *CN* and *CL* ($M=5.35$, $CI [4.15 \ 6.20]$) and conditions *EL* and *CN* ($M=-5.23$, $CI [-6.29 \ -4.00]$). In all cases, the effect size for change of feature was smaller than that for change in knowledge of object location.

Findings: Object location knowledge produced smaller and similar F_c for both feature conditions. When object location was unknown, F_c were higher, particularly when CCT was the mismatched feature. Thus, we reject *H1* and accept the alternative hypothesis that both knowledge of object location and changes in mismatched feature (when object location is not known a priori) impact F_c .

4.1.2 Fixation Duration (F_d)

Results: Permutation testing found no significant F_d differences between conditions *EN* and *CN* ($p = 0.33$) (see Fig.3), however F_d values for each of conditions *EL* and *CL* were significantly different from every other condition at $p < .05$ (see Fig.4). The largest effect size was observed for mean differences between conditions *CN* and *CL* ($M=-.19$, $CI [-.23 \ -.15]$). A smaller effect size of this type was also present across the location condition for exposure: conditions *EN* and *EL* ($M=-.07$, $CI [-.11 \ -.03]$).

Findings: When object location was known, F_d for both mismatched features were significantly longer than when object location was unknown. F_d were also significantly longer for CCT compared to exposure when object location was known. Given this evidence, we reject *H2* and accept the alternative hypothesis that both observer knowledge of object location, and difference in mismatched feature have an effect on F_d .

4.1.3 Time to First Fixation on Object ($TFFO$)

Results: Fisher's permutation test found no significant differences between any of the conditions.

Findings: As no significant median $TFFO$ differences were found between conditions, the null hypothesis *H3* that varying the mismatched object feature and knowledge of object location/identity has no effect on $TFFO$ cannot be rejected. This suggests that for the parameters used in our analysis, neither varying the mismatched object feature, nor revealing object location cause observers to attend to the object any sooner. Based on this we cannot reject *H3* that neither varying the mismatched feature between exposure and CCT, nor changing object location information has an impact on $TFFO$.

4.1.4 Duration of First Fixation on Object ($DFFO$)

Results: Fisher's permutation test revealed significant differences between the medians of each of the *EL* and *CL* conditions (see Fig.3) and every other condition at $p < .05$. As in the case of F_d , no significant difference was found between conditions *EN* and *CN* ($p = 0.56$).

Again, the largest effect size was present when comparing between medians of conditions *CN* and *CL* ($M=-.59$, $CI [-.68 \ -.45]$), as well as *EN* and *CL* ($M=-.57$, $CI [-.66 \ -.44]$). Notably, a more pronounced effect size between conditions *EN* and *EL* ($M=-.24$, $CI [-.31 \ -.09]$) can be noted, compared to the F_d metric.

Findings: $DFFO$ were longer when object location was known, particularly in the case of CCT. When object location was unknown, no significant differences in $DFFO$ were found. Group differences for this metric mirror the order of those for F_d . Thus, hypothesis *H4* is rejected and the alternative hypothesis that both revealing object location to observers, as well as changing the mismatched feature from exposure to CCT has an effect on $DFFO$.

4.2 Fixation Map Metrics

4.2.1 Inter-Observer Consistency (IOC)

Results: Fisher's permutation test indicated significant mean IOC differences between conditions *EL* and *CL*, *EL* and *CN*, as well as *EN* and *CN* at $p < .05$ (see Fig. 4). However, their effect sizes are small, with the largest being between conditions *EL* and *CL* ($M=-.03$, $CI [-.04 \ -.01]$).

Findings: Fixations of observers in the group assessing CCT off-sets of a known object received highest IOC scores. This could be due to their increased focus on the target object both in spatial and temporal terms, as indicated by other metrics discussed here. Revealing object location did not significantly change consistency for either feature. Observers in each condition show a similar degree of spatial consistency. As significant differences between conditions were found, we reject *H5* and accept the alternative hypothesis that varying both the mismatched feature and observer knowledge of object location have an effect on IOC .

4.2.2 Area of Interest Similarity ($AOIS$)

Results: Bootstrapped mean $AOIS$ values along with their 95% CIs obtained for comparisons of joint fixation maps across pairs of conditions are visualised in Figure 5. Highest similarity values here were obtained between pairs *EN* and *CN* ($M=.81$, $CI [.78 \ .82]$), as well as *EL* and *CL* ($M=.82$, $CI [.81 \ .84]$). A Fisher's permutation test found significant differences between the mean similarities of each of these two condition combinations and every other condition combination at $p < .05$. No other significant differences were found.

Findings: The $AOIS$ results indicate that object location knowledge had a higher effect on attended image regions than mismatched feature. The similarity between fixation distributions for conditions involving both features and one location condition was consistently higher than the similarity between conditions involving one feature and a change in the object location condition. Based on this evidence we reject *H6* and accept the alternative hypothesis that object location knowledge changes $AOIS$.

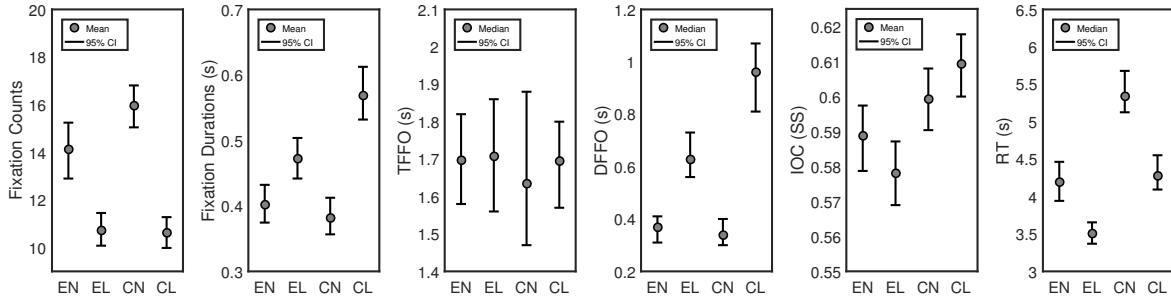


Figure 3: Bootstrapped means/medians and their 95% confidence intervals for the evaluated metrics for test images under the four experimental groups - exposure, no location (EN); exposure, location (EL); CCT, no location (CN); CCT, location (CL). From left: fixation counts (F_c), fixation durations (F_d), time to first fixation on target object ($TFFO$), duration of first fixation on target object ($DFFO$), inter-observer consistency using similarity score (IOC_{ss})

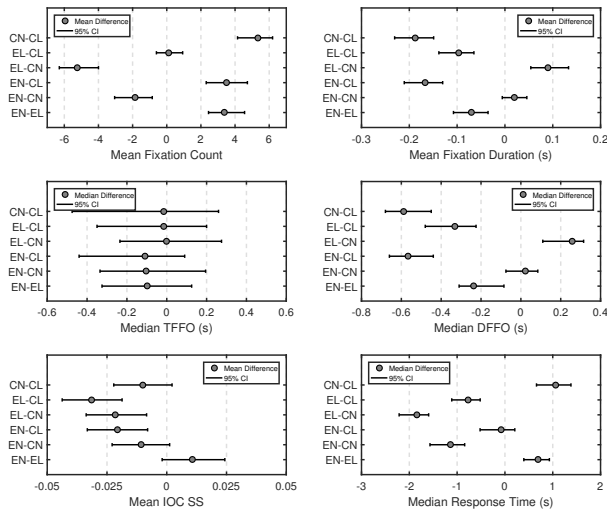


Figure 4: Bootstrapped comparisons of group mean/median differences for each of the evaluated metrics.

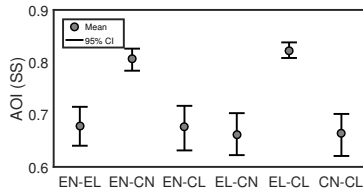


Figure 5: Test image AOIS for each pairwise combination of conditions.

4.3 Realism Responses

4.3.1 Response Times (RT)

Results: Bootstrapped median RT values along with their 95% CIs obtained for each condition are visualised in Figure 3. A Fisher's permutation test did not find significant median RT differences between these two conditions ($p = .63$). Significant differences were found for every other combination of conditions at $p < .05$. Bootstrapped means, their CIs and effect sizes for these group differences are visualised in Figure 4. The largest effect sizes were observed between conditions EL and CN ($M = -1.84$, $CI [-2.18 - 1.58]$), EN and CN ($M = -1.15$, $CI [-1.57 - .85]$), as well as CN and CL ($M = 1.06$, $CI [.67 - 1.37]$).

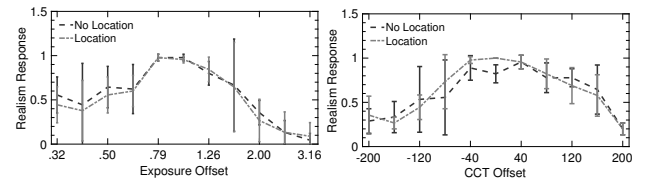


Figure 6: Realism responses averaged for feature offset values across image sets for exposure (left) and CCT (right). Line styles indicate object location conditions. Error bars indicate standard deviation of mean realism ratings for each offset level.

Findings: Median response times for CCT were consistently higher compared to exposure, within each of the location conditions and significantly lower across the location condition for each of the features. The latter difference indicates that additional information contributed to faster decisions regarding realism, the former is in alignment with other fixation metrics, suggesting that CCT mismatches may have been more difficult to detect and/or judge by observers. Based on this evidence, we accept $H7$ stating that revealing object location to observers reduces RT . We note a significant increase in RT when the mismatched feature is changed from exposure to CCT in a given object location knowledge condition.

4.3.2 Realism Ratings

Results: Realism scores are shown in Figure 6. Despite the between-subjects design, observers' realism responses were highly correlated across location conditions, particularly between conditions EN and EL ($r = .91$, $p < .05$) and CN and CL ($r = .81$, $p < .05$). Between-feature correlations for conditions EN and CN ($r = .57$, $p < .05$), as well as EL and CL ($r = .63$, $p < .05$) were weaker. This supports $H8$ that realism ratings should not change across location conditions.

Findings: While revealing the identity and location of mismatched objects affected almost every gaze metric under test, it did not have a large impact on observers' realism responses, which remained highly correlated for both features across location conditions. This suggests that once the target object was located by observers, visually similar trends in realism judgements were made for each mismatched feature (see Figure 6). Expectedly, correlation between realism responses was lower when comparing exposure with CCT. This supports the explanation that CCT offsets were more difficult to judge, compared to exposure under the conditions of our study. This evidence supports $H8$ that realism responses should not change across the location condition.

5 CONCLUSIONS

This paper has presented a first quantitative study into gaze patterns of observers analysing the visual realism of image composites and the impact of object location and identity knowledge in this context. A series of gaze metrics have been presented and evaluated in a 2×2 factorial study involving 60 observers. We found that the metrics selected provide insight into the visual processes involved in observers analysing visual realism/quality. Furthermore they allow for differentiation between spatial and temporal gaze properties, which proves useful for studying the process of composite analysis.

Under the experimental conditions outlined, our results point towards several conclusions:

1. Observers focus primarily on the object in question and much less on other scene regions when performing assessment.
2. Knowledge of object location impacts the process of analysis (i.e. fixation metrics), but has little effect on the outcome (i.e. the final realism ratings).
3. Revealing/withholding object location has a larger impact on spatial fixation allocation than varying the mismatched feature between exposure and CCT. Along with minimal differences in inter-observer consistency this suggests the possibility of common spatial strategies for CCT and exposure.
4. Significant differences in fixation metrics, particularly fixation durations and counts, suggest the extraction of task-relevant information for CCT mismatches may be more difficult than for exposure. Significant differences in response times reinforce this conclusion.

Our results indicate that the eye-tracking paradigm is a reliable way to improve our understanding of how observers perform visual realism judgements. Parallels can also be drawn between the results of our work and similar previous studies. For example, as the location of mismatched objects was indicated to observers, their fixation durations increased, similarly to [15] and in accordance with predictions of [9]. Greater concentrations of fixations were noted when observers had prior knowledge of object locations, as in [28].

REFERENCES

- [1] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.
- [2] P. Cavanagh. The artist as neuroscientist. *Nature*, 434(7031):301–307, 2005.
- [3] A. Dolhasz, I. Williams, and M. Frutos-Pascual. Measuring observer response to object-scene disparity in composites. In *Proceedings of the 15th International Symposium on Mixed and Augmented Reality (ISMAR)*, Merida, Mexico, 2016.
- [4] A. T. Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, 34(4):455–470, 2002.
- [5] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
- [6] M. Elhelw, M. Nicolaou, A. Chung, G.-Z. Yang, and M. S. Atkins. A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Transactions on Applied Perception (TAP)*, 5(1):3, 2008.
- [7] U. Engelke, R. Pepion, P. Le Callet, and H.-J. Zepernick. Linking distortion perception and visual saliency in h. 264/avc coded video containing packet loss. In *Visual Communications and Image Processing 2010*, pages 774406–774406. International Society for Optics and Photonics, 2010.
- [8] A. Gegenfurtner, E. Lehtinen, and R. Säljö. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, 23(4):523–552, 2011.
- [9] J. M. Henderson and A. Hollingworth. Eye movements during scene viewing: An overview. *Eye guidance in reading and scene perception*, 11:269–293, 1998.
- [10] J. M. Henderson and A. Hollingworth. High-level scene perception. *Annual review of psychology*, 50(1):243–271, 1999.
- [11] R. Jacob and K. S. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, 2(3):4, 2003.
- [12] S. D. König and E. A. Buffalo. A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds. *Journal of neuroscience methods*, 227:121–131, 2014.
- [13] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013.
- [14] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In *2007 IEEE ICIP*, volume 2, pages II–169. IEEE, 2007.
- [15] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, and A. Tirel. Task impact on the visual attention in subjective image quality assessment. In *Signal Processing Conference, 2006 14th European*, pages 1–5. IEEE, 2006.
- [16] A. Odén and H. Wedel. Arguments for fisher’s permutation test. *The Annals of Statistics*, pages 518–520, 1975.
- [17] J. L. Orquin and S. M. Loose. Attention and choice: A review on eye movements in decision making. *Acta psychologica*, 144(1):190–206, 2013.
- [18] A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.
- [19] U. Rajashekar, L. K. Cormack, and A. C. Bovik. Point-of-gaze analysis reveals visual search strategies. In *Electronic Imaging 2004*, pages 296–306. International Society for Optics and Photonics, 2004.
- [20] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.
- [21] I. Recommendation. 500-11, methodology for the subjective assessment of the quality of television pictures, recommendation itu-r bt. 500-11. *ITU Telecom. Standardization Sector of ITU*, 2002.
- [22] J. Redi, H. Liu, R. Zunino, and I. Heynderickx. Interactions of visual attention and quality perception. In *IS&T/SPIE Electronic Imaging*, pages 78650S–78650S. International Society for Optics and Photonics, 2011.
- [23] E. D. Reichle, A. Pollatsek, D. L. Fisher, and K. Rayner. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125, 1998.
- [24] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1153–1160, 2013.
- [25] A. R. Robertson. Computation of correlated color temperature and distribution temperature. *JOSA*, 58(11):1528–1535, 1968.
- [26] Tobii. Tobii x1 light eye tracker, 2012.
- [27] M. V. Venkatesh and S. C. Sen-ching. Eye tracking based perceptual image inpainting quality analysis. In *2010 IEEE International Conference on Image Processing*, pages 1109–1112. IEEE, 2010.
- [28] C. T. Vu, E. C. Larson, and D. M. Chandler. Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. In *Image Analysis and Interpretation, 2008. SSIAI 2008. IEEE Southwest Symposium on*, pages 73–76. IEEE, 2008.
- [29] J. M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.
- [30] J. M. Wolfe. Visual attention. *Seeing*, 2:335–386, 2000.
- [31] S. Xue, A. Agarwala, J. Dorsey, and H. Rushmeier. Understanding and improving the realism of image composites. *ACM TOG*, 31(4):84, 2012.
- [32] A. L. Yarbus. *Eye movements during perception of complex objects*. Springer, 1967.
- [33] W. Zangemeister, K. Sherman, and L. Stark. Evidence for a global scanpath strategy in viewing abstract compared with realistic images. *Neuropsychologia*, 33(8):1009–1025, 1995.

Learning to Observe: Approximating Human Perceptual Thresholds for Detection of Suprathreshold Image Transformations

Alan Dolhasz, Carlo Harvey, Ian Williams
Digital Media Technology Lab, Birmingham City University
{alan.dolhasz, carlo.harvey, ian.williams}@bcu.ac.uk
<https://github.com/dmt-lab/learning-to-observe>

Abstract

Many tasks in computer vision are often calibrated and evaluated relative to human perception. In this paper, we propose to directly approximate the perceptual function performed by human observers completing a visual detection task. Specifically, we present a novel methodology for learning to detect image transformations visible to human observers through approximating perceptual thresholds. To do this, we carry out a subjective two-alternative forced-choice study to estimate perceptual thresholds of human observers detecting local exposure shifts in images. We then leverage transformation equivariant representation learning to overcome issues of limited perceptual data. This representation is then used to train a dense convolutional classifier capable of detecting local suprathreshold exposure shifts - a distortion common to image composites. In this context, our model can approximate perceptual thresholds with an average error of 0.1148 exposure stops between empirical and predicted thresholds. It can also be trained to detect a range of different local transformations.

1. Introduction

Human observers are the target audience for image content and thus the ultimate judges of image quality, which is often measured with reference to opinions of humans and various *local* and *global* distortions and inconsistencies perceptible to them. These distortions can arise as side-effects of image acquisition, compression, transmission, compositing, and post-processing. Understanding and modeling how humans detect and process distortions to arrive at subjective quality scores underpin image quality assessment (IQA) research. Many attempts have been made at modeling the sensitivity of the human visual system (HVS) to certain types of distortions for applications primarily in IQA [13, 8, 24, 54, 16, 23] and saliency modeling [51, 55, 29, 19], where detection of relevant and perceptu-

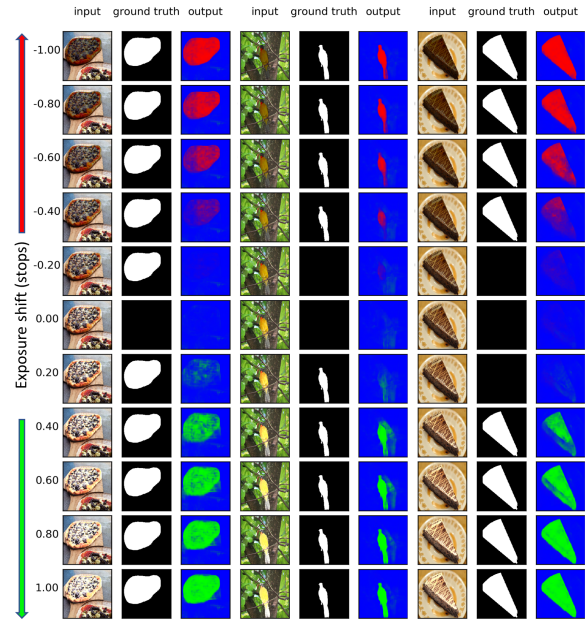


Figure 1. Performance of our model illustrated for three input images and 11 levels of exposure transformation. The left columns show input images with applied exposure transformations and the magnitude of this transformation expressed on a \log_2 scale. Middle columns show ground truth from our subjective experiments and rightmost columns show output of our model, where red and green regions indicate detected negative and positive suprathreshold exposure transformations, while blue regions indicate no suprathreshold transformations.

ally suprathreshold features is key to the approximation of human performance. However, many of these approaches are limited in their generalizability, efficiency or transferability. Alternative approaches based on signal fidelity [45], statistical measures [46] and deep learning models [7, 50] were also developed as a way to address such limitations.

Human sensitivity to physical stimuli is measured using

psychophysics [17] and often represented using psychometric functions, which describe observer performance as a function of stimulus intensity [2]. This method is effective when stimuli are simple, but is difficult to generalize to more complex stimuli, such as natural images. This is largely due to the vast amount of variation in the set of natural images and the corresponding number of trials required to measure observer performance across sufficiently many images and stimulus intensities. In subjective image evaluation, the quality score can be seen as a result of applying an *observer function* to an input image. This function can be summarized as detection of visible distortions, their implicit pooling, and mapping to a point on a given quality scale [20]. This is further influenced by task, image content, and allocation of attention [33]. Recent work has made significant progress in approximating this entire process in the context of IQA using deep convolutional neural networks (DCNN) [7, 50]. However, these approaches are mostly limited to a fixed set of low level, globally-distributed artifacts available in public IQA datasets, such as LIVE [47] containing 5 types of distortions, or TID2013 with 24 types and 5 magnitude levels each [39]. This limits the generalizability, particularly for applications where the type and number of possible distortions vary significantly, or where the distortions are context-dependent and only present in a local region of the image, such as image compositing. The creation of such datasets is a costly and time-consuming process, due to the need for human observers. Approximation of this observer function - detecting visible inconsistencies of an arbitrary type - would allow for application in many areas related to IQA, including composite quality assessment, manipulation detection, and image restoration.

In this work, we propose a DCNN-based methodology to approximate this observer function and validate our method with respect to a specific local distortion common to image composites - *local exposure inconsistencies* associated with an image region occupied by an object. We achieve this by learning a mapping between images affected by this distortion and corresponding points on an empirical psychometric function, estimated with respect to this distortion type. Viewing image distortions as transformations allows use of unsupervised methods for learning relevant features. Our approach can be applied to a range of problems where distortions visible to humans need to be localized in an image, such as IQA or composite quality assessment, even when little subjective data is available. Our contributions are:

- A novel method for detecting effects of local image transformations based on perceptual data and unsupervised pre-training
- A model trained using this method to detect local exposure shifts
- A dataset of images with corresponding empirical subjective perceptual thresholds from our experiments

2. Related Work

2.1. Human Perception

The HVS displays different levels of sensitivity to various distortions and inconsistencies in images, detecting some readily [5], while disregarding others completely [36, 9]. Detection of inconsistencies in lower-level properties of images depends largely on fundamental characteristics of the HVS, such as contrast sensitivity [2], luminance adaptation, and masking [38]. These characteristics describe how immediate context, such as differences in background luminance, spatial frequency, and presence of texture, influence the visibility of different image artifacts. For example, distortions such as noise or quantization, are much easier to notice on a textureless background, compared to a textured one. The amount of change in stimulus required for an observer to reliably notice a difference is referred to as the a just-noticeable difference (JND) or difference limen. JNDs have been used extensively to model human perceptual sensitivity in tasks such as blur detection [48], visual attribute differences [60], perceptual metrics [63], or 3D model attribute similarities [12]. Observer sensitivity is further modulated by the allocation of visual attention [35, 30], particularly for localized distortions, such as those in image composites [14].

2.2. Psychometric Functions

Observers assessing image quality base their judgments on visual evidence, such as visible artifacts or distortions [52]. Human performance in detection and discrimination tasks is commonly modeled using psychometric functions [49, 44, 21, 53, 34]. The psychometric function describes a relationship between observer performance and an independent variable, often describing a stimulus level or physical quantity [57]. It is defined as

$$\Psi(x; \theta) = \gamma + (1 - \gamma)f(x; \alpha, \beta) \quad (1)$$

where θ refers to the set of parameters: γ (guess rate) defines the lower bound of the function corresponding to chance performance, while $f(x; \alpha, \beta)$ defines a sigmoidal function parametrized by α - its location and β - its slope. Observer performance for a given stimulus x is represented by the output of Ψ denoted as $y = \Psi(x; \theta)$. The threshold of a perceptual function can thus be defined as the stimulus level x_t which yields a particular probability of stimulus detection y_t , such that $x_t = \Psi^{-1}(y_t)$. In practice, psychometric functions are commonly estimated using adaptive sampling procedures, such as QUEST [56], which limit the number of required trials by sampling stimuli with the highest probability of lying at the threshold.

2.3. Saliency & Semantic Segmentation

Our work is related to both salient object detection (SOD) and semantic segmentation (SS), both of which seek to assign class membership of individual pixels based on local contextual information. SS assigns a single semantic object class to each pixel of an input image [31]. SOD aims to segment the most salient object in an image, based on its low-level image-based features, often measured against human performance [6]. Image-to-Image neural networks have become popular tools in these domains, underpinning many state-of-the-art CNN architectures such as fully-convolutional networks (FCNs) [11], U-nets [42], adversarial approaches, such as Pix2Pix [18] and many variations thereof. These approaches emphasise the importance of multi-scale features [25], as well as spatial resolution preservation through dilated convolution and skip connections [61, 10].

2.4. Unsupervised & Semi-Supervised Learning

Supervised learning approaches, such as those in Section 2.3, require large amounts of labeled data, which can necessitate a significant time effort. For perceptually-constrained tasks, this overhead is multiplied, due to the requirement for larger observer samples and more replications, compared to Likert-style subjective opinion studies. Conversely, unsupervised learning techniques do not require manually-labeled data to learn. Thus, this paradigm is attractive for our application, as we can exploit unlabelled data to learn the features describing a given transformation and then use a smaller, labeled perceptual dataset to fine-tune these features to the empirical perceptual data.

Some approaches, such as representation learning [3], relax the requirement for labeled data through the use of auto-encoders (AEs) and generative adversarial networks (GANs). AEs learn compressed representations of data by attempting to reconstruct it through a feature bottleneck. Representations learned by AEs tend to encode salient features of the data they are conditioned on, which in turn can act as a task-specific feature extractor for supervised tasks [1]. On the other hand, GANs adopt an adversarial training regime, where a generator and discriminator are jointly trained. E.g. the generator can be tasked with generating a sufficiently realistic image, such that the discriminator classifies it as real. In turn, the discriminator is tasked with separating generated images from real ones [40]. Zhang et al. (2019) showed that the performance of supervised classifiers can be improved by using an Auto-Encoding Transformations paradigm. They propose to learn transformation equivariant representations (TERs), which encode transformations applied to the input [62]. This reduces the need for data augmentation and forces the encoder to learn a better representation of the input data, which encodes visual structures well, invariant of the transformation of the input.

We adapt this approach to detecting local transformations within an image, which forms the foundation of our proposed methodology.

3. Method

In this section, we elaborate on our proposed approach and detail our model design and rationale. We summarize our methodology, including the formulation of distortions as transformations, use of empirical perceptual thresholds as decision boundaries, collection of empirical psychometric data, training dataset preparation, and both stages of our training procedure.

3.1. Distortions as Transformations

Many distortions affecting image quality can be seen as transformations applied to the original, uncorrupted image as a side-effect of some processes such as transmission, compositing, or compression. This is conceptually similar to the intuition behind denoising autoencoders [4]. Denoising autoencoders learn a low-dimensional manifold near which training data concentrate. They also implicitly learn a function projecting corrupted images \tilde{I} , affected by a corruption process and lying near the manifold of uncorrupted images, back onto this manifold. This conceptualization allows for the generation of large amounts of training data from a small set of undistorted images, by applying various transformations. We focus on a single transformation: local exposure shifts. This corresponds to the scaling of luminance by a constant, applied to a region within image I corresponding to an object and defined by a binary mask M . This is performed on the luminance channel of the perceptually-uniform *Lab* colorspace [41]. We motivate this choice as follows: observers are reliable at detecting such low-level image distortions [15]; exposure distortions represent common mismatches present in image composites, which are a motivating application of our research [59]; this type of transformation is computationally inexpensive to apply, allowing for gains in training efficiency.

3.2. Perceptual Thresholds as Decision Boundaries

In the context of image distortions and assuming controlled viewing conditions, a psychometric function can be seen as the result of an observer process operating on a range of input data. Given an unprocessed image I , object mask M , observer function O and \tilde{I}_x a corrupted version of I resulting from a local transformation $T(I, M, x)$, the empirical psychometric function can be interpreted as a result of applying the observer function to \tilde{I} for all values of x . The observer function O thus represents the perceptual process performed by an observer, which maps an input stimulus \tilde{I}_x to a point on the psychometric function. Accordingly, detecting suprathreshold transformations in an image can be defined as applying the observer model to classify each

pixel based on the existence of the effects of a suprathreshold transformation. This requires **a)** a psychometric function describing observer performance with respect to the magnitude of the transformation and specific image stimulus, **b)** contextual information about the scene and appearance of objects within it, from which information about the existence of local distortions can be derived and **c)** an appropriate feature representation, equivariant to the transformation in the training data. Consequently, our problem can be defined as a *pixel-wise classification* of an image, where each pixel is assigned one of three classes c , whose decision boundaries are defined by the thresholds x_{t-} and x_{t+} of the two psychometric functions estimated for a given image, with respect to the parameter x of the transformation generating the stimuli \tilde{I} :

$$c = \begin{cases} 0, & \text{if } x < x_{t-} \\ 1, & \text{if } x > x_{t+} \\ 2, & \text{otherwise} \end{cases} \quad (2)$$

Here, x_t is the value of the transformation parameter for which the probability of detection exceeds threshold t , set to 0.75, corresponding to the JND in 2AFC tasks. This is the midpoint between perfect (100%) and chance (50% for 2AFC task) performance [57]. As we capture two psychometric functions per image, one corresponding to decreasing the pixel intensity (x_{t-}) and one for increasing it (x_{t+}), their two thresholds separate the parameter space x into three regions (Fig. 2d).

3.3. Psychometric Function Estimation

To estimate image-wise empirical psychometric functions with respect to our transformation, we design a 2AFC study using a dataset of natural images with segmented objects, where the segmentation is defined by a binary mask. Following the approach of [15], we systematically apply transformations with different values of x to the segmented object. We display the *original* (I) and *transformed* (\tilde{I}) images side by side in random order and ask observers to identify I correctly. We repeat this for multiple values of x and fit Weibull psychometric functions to each observer's responses for each image. To extract the thresholds, we estimate the parameter values x_{t-} and x_{t+} corresponding to a performance level of y_t for negative and positive exposure shifts, respectively. We then bootstrap mean thresholds across all observers who viewed the same image. We detail the stages of this process in the remainder of this section.

3.3.1 Experiment Design

All experiments are performed under controlled laboratory conditions, following the ITU BT-500 recommendation [20]. We use an Apple Cinema HD 23" monitor, calibrated to sRGB colorspace using an X-Rite i1Display Pro

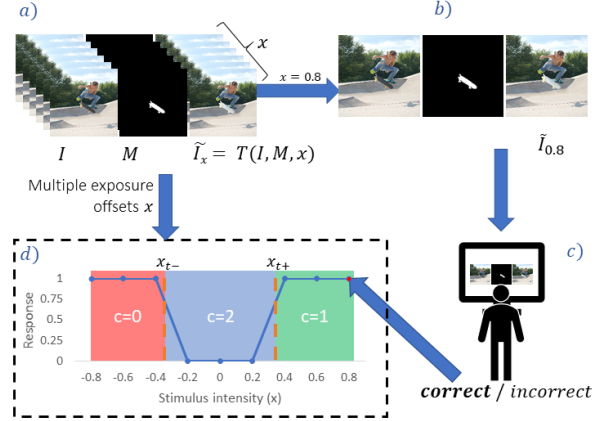


Figure 2. Illustration of the 2AFC procedure used in our experiments. **a)** For a given image I and object mask M we generate images \tilde{I} with different exposure offsets based on the sampled value of x . **b)** Example stimulus displayed to an observer. **c)** Observer correctly identifies I and \tilde{I} for $x = 0.8$. **d)** Observer response added to their previous responses for different sampled values of x . Symbols x_{t-} and x_{t+} , illustrated with orange dashed lines, indicate the location of the threshold after performing psychometric function fitting.

display calibration device. Observers are positioned 65cm away from the display. To mitigate the confounding impact of visual search on the task, particularly when differences between the images are minimal, we explicitly indicate the transformed region in the image by displaying the binary mask corresponding to the object, following [14]. To minimize the number of experimental trials we leverage the QUEST adaptive sampling procedure [56], using the implementation from the PsychoPy 2 library [37].

3.3.2 Observers & Stimuli

We recruit $N = 120$ naive observers, with a mean age of 31 ($SD = 11.85$), 44 of whom are female and randomly assign them to 20 groups. Observers are screened for normal vision before participating in the experiment. Our stimuli dataset consists of 300 8-bit images with corresponding object masks, randomly sampled from the LabelMe [43] and SUN [58] datasets. These images are then evenly distributed across the observer groups. Each group views 15 unique images from the dataset.

3.3.3 Task & Experimental Procedure

In the experimental session, each observer performs repeated 2AFC trials for each of the 15 base images in their allocated image sample, viewing at least 20 different variations of each base image. Observers first complete 20 trials using a calibrating image, results for which are discarded.

In each trial observers are shown 2 images: the original image I and a transformed version of the original image \tilde{I}_x , the result of exposure transformation $T(I, M, x)$ of magnitude x . A segmentation mask M is also displayed indicating the target object. These images are displayed at the same time and remain on-screen for 5 seconds. The order of I and \tilde{I} is randomized every trial. Observers are instructed to correctly indicate I by clicking a corresponding button. After each response, a new value of x is sampled by the QUEST procedure [56], and the process is repeated 20 times.

3.3.4 Perceptual Threshold Estimation

For each observer-image combination, we collect binary responses y with corresponding stimulus intensities x . We use the PsychoPy library [37] to fit a Weibull cumulative distribution function to this data, given by

$$y = 1 - (1 - \gamma)e^{-\left(\frac{kx}{t}\right)^\beta} \quad (3)$$

and

$$k = -\log\left(\frac{1 - \alpha}{1 - \gamma}\right)^{\frac{1}{\beta}} \quad (4)$$

where x is the stimulus intensity, y is the proportion of correct responses, γ is the performance level expected at chance, equal to 0.5 for 2AFC tasks, α is the performance level defining the threshold (set to 0.75, corresponding to the JND for 2AFC), β is the slope of the function and t is the threshold. Once we extract the threshold of this function, we pool the threshold values across observers for that image and bootstrap the mean of these thresholds, using 1000 bootstrap samples. We obtain two generalized perceptual thresholds: x_{t-} and x_{t+} for each image in our dataset.

3.4. Transformation Equivariant Representation Learning (AET)

While object classifiers, such as models trained on ImageNet, aim to achieve invariance to changes in object brightness, our task explicitly uses these features to assign classes to output pixels. Thus, transfer learning with an object classifier/detector is unsuitable for addressing overfitting with our small dataset. Instead, we propose to first learn a task-specific TER in an unsupervised manner, adopting the AET approach of Zhang et al. [62], who encode a TER by training to predict transformation parameters that describe a transformation between two inputs. Analogously, we wish to encode a representation that is invariant to a particular transformation type: local exposure shifts.

3.4.1 AET: Network Architecture

We can train a convolutional autoencoder to predict the parameter of a local exposure shift applied to the input, by

mapping images containing local exposure shifts to masks indicating their pixel-wise magnitude. To achieve this, we develop an AET model based on the VGG16. We first convert the VGG16 to a fully convolutional network [31]. Due to the importance of contextual and multiscale information to our task, we attach a multiscale extension, as proposed in [26]. This introduces skip connections to the model, taking outputs after each max pooling layer in the VGG16 and passing each through an additional convolutional branch before concatenating the output of all branches. Each branch consists of 3 convolutional blocks. The first block contains a 3×3 , 128-channel convolutional layer with a stride setting dependant on the scale of the input. This is 4, 2, 1, 1 respectively for inputs from the first 4 max pooling layers, causing all multiscale branches outputting feature maps of equal resolution. This layer is followed by a batch normalization layer and a ReLU activation. The following two blocks contain 1×1 convolutional layers with a stride of 1, with 128 and 3 channels respectively. They are each followed by batch normalization and a ReLU activation. To output masks of equal resolution to the input images, we add a convolutional decoder to the output of the multiscale concatenation layer in our model. It consists of 3 blocks, each block containing a $2 \times$ upsampling layer, followed by two sets of convolution, batch normalization, and ReLU layers. The first convolution in the block uses 3×3 kernels, while the second uses 1×1 kernels. See Figure 3 for a detailed overview.

Using this architecture, we design an AET model which shares the weights of the network between two image inputs, I and \tilde{I}_x (Fig. 4). Activations for both inputs are concatenated and fed to a final convolutional layer. As our transformation can be expressed by a single scalar the final layer of our AET is a 3×3 convolutional layer with a linear activation, which outputs masks with resolution equal to the input image, with a single value expressing the predicted exposure shift for each pixel. This way we can train our model to approximate pixel-wise transformations applied to an input image.

3.4.2 AET: Training Data Generation

To train the AET in an unsupervised manner, we learn a mapping between input images \tilde{I} and output masks $Y = xM$, which encode the parameter of the transformation applied to the input. \tilde{I} contains an exposure shift applied within the region defined by M . Each pixel in Y contains the value of the exposure shift x applied to the corresponding pixel in \tilde{I} . This is x wherever $M = 1$ and 0 elsewhere (Fig. 4). During training, we dynamically sample images I and corresponding masks M from the MSCOCO dataset [28]. As some images in MSCOCO contain multiple masks, we randomly select one of them, provided its area is larger

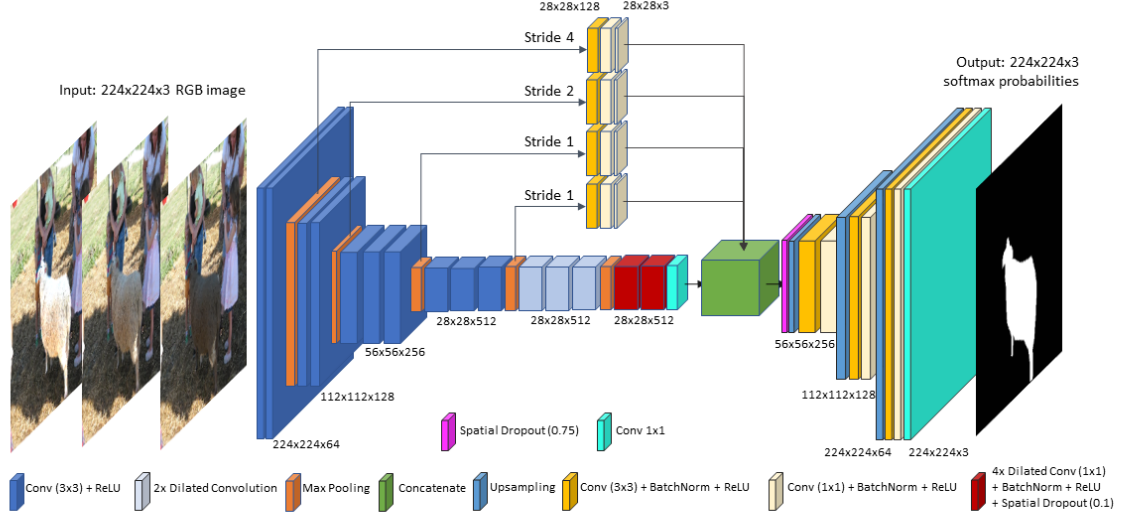


Figure 3. Architecture of our VGG16-based convolutional autoencoder used in the perceptual threshold learning task. The network is based on a FCN adaptation of the VGG16. See Section 3.4 for a detailed description of the architecture.

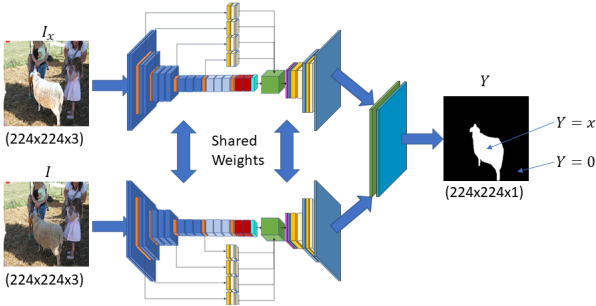


Figure 4. Unsupervised AET architecture consisting of a VGG16-based convolutional autoencoder with weights shared across two inputs. Activations for both inputs are then concatenated and fed to a final convolutional layer with a single channel output. The output masks encode the parameter of the transformation for each pixel.

than 1%. We then apply exposure shifts by sampling the transformation parameter x and scaling the luminance channel of I after conversion to Lab colorspace:

$$\tilde{I}_L = 2^x I_L \odot M + I_L \odot (1 - M) \quad (5)$$

where x is a scalar sampled from a base-2 log-uniform distribution spanning $(\log_2(0.1), \log_2(10))$, I_L is the luminance channel of the original image I after conversion from RGB to Lab colorspace, M is the alpha mask and \odot is the Hadamard product. We clip the pixel values of processed image to the range $(0.0, 1.0)$, convert back to RGB , rescale to 0.0 mean and unit variance, reshape images to $(224, 224, 3)$ and feed both I and \tilde{I} to the two inputs of the AET (as in Fig. 4). The output of the network is a mask \hat{Y} approximating the parameter of the transformation at each

pixel of the input image.

3.4.3 AET: Objective & Optimizer Details

We train our model using the Adam optimizer [22]. We use default values for all parameters, aside from the learning rate, which is controlled using a cosine annealing schedule [32]. The minimum and maximum learning rate in the annealing schedule are set to $1e-6$ and $1e-4$, respectively. The learning rate cycles between these values over 5 epochs, after which the maximum learning rate is reduced to 90% of its value, and the cycle is repeated for $1.5\times$ as many epochs. We train the AET for 90 epochs, minimizing the mean squared error (MSE) loss between \hat{Y} and Y . We use the model with the lowest validation error as the backbone for the Perceptual Threshold Classifier.

3.5. Perceptual Threshold Classifier (PTC)

3.5.1 PTC: Network Architecture

To detect perceptually suprathreshold transformations in images, we utilize the pre-trained AET architecture described in Section 3.4, extract the encoder and decoder shown in Figure 3 and replace the final single-channel convolutional layer of the decoder with a spatial dropout layer with a dropout probability of 75%, followed by a 3-channel convolutional layer with a softmax activation.

3.5.2 PTC: Training Data Generation

Using thresholds obtained in our experiments, we devise a data generation method which dynamically applies random exposure transformations to the images used in our 2AFC

experiment and generates corresponding categorical masks, based on whether the parameter of the transformation x exceeds one of the two empirical thresholds estimated for a given image. When x exceeds a threshold, any pixels affected by this suprathreshold transformation are assigned $c = 0$ (negative suprathreshold exposure shift) or $c = 1$ (positive suprathreshold exposure shift), following Equation 2. The last channel of the target image corresponding to $c = 2$ is conceptually similar to the *background* class in semantic segmentation models. It indicates pixels that do not belong to any of the foreground classes. In our case, these are pixels unaffected by a suprathreshold transformation. We use a 90%-10% training/validation split. The shape of the target mask is (224, 224, 3), containing one channel per class. During training, we use a data generator constrained to ensure a balanced class distribution in each minibatch. Specifically, for each batch we sample x from three random distributions whose ranges are defined by the perceptual thresholds for a given image:

$$x \in \mathbb{R} : \begin{cases} (\log_2(0.1), x_{t-}), & \text{if } x < x_{t-} \\ (x_{t+}, \log_2(10)), & \text{if } x > x_{t+} \\ [x_{t-}, x_{t+}], & \text{otherwise} \end{cases} \quad (6)$$

The distribution for $c = 2$ is log-uniform, whereas the distributions for classes 0 and 1 are exponential distributions biased towards values of x lying close to the thresholds x_{t-} and x_{t+} respectively. These three values of x are then used to create three processed images and corresponding target masks Y , one for each class. For larger batch sizes we simply sample multiple images for each class. To improve generalization, we apply image augmentation, limiting to zooming, rotation, and cropping in order not to affect relative pixel intensities. We perform horizontal and vertical flipping with 50% probability, as well as random scaling and cropping in the range 110-150% and with 50% probability.

3.5.3 PTC: Objective & Optimizer Details

We follow the optimization approach from Section 3.4.3 with minor changes. Firstly, we select a loss function appropriate for pixel-wise classification with an imbalanced dataset. In most images in our dataset the background class occupies more pixels than either of the suprathreshold classes, we handle this imbalance by reducing the contribution of easy classification examples to the loss using focal loss [27]. We also experiment with freezing different sections of our backbone network in order to maximize generalizability. We train our models with a batch size of 12 until convergence using early stopping to cease training when no improvement in validation loss is seen for 400 epochs. For evaluation, we select the model which maximizes the validation mean intersection-over-union measure.

4. Results & Discussion

4.1. Perceptual Threshold Estimation

In our 2AFC study, we obtained a total of 41725 unique responses, with an average of 23.14 responses per observer per image. Observers took on average 2.65s per response. A total of 590 mean thresholds for 295 images were calculated after fitting psychometric functions, bootstrapping and removing outlier thresholds beyond 3 standard deviations (Fig. 5). The means of the resulting threshold distributions were $x_{t-} = -0.2478$ and $x_{t+} = 0.2280$ for negative and positive thresholds respectively. On average, perceptual thresholds were lower for highly-textured and bright objects. We found significant correlations between the mean luminance of target objects and corresponding mean thresholds. For negative offsets the Pearson product-moment correlation coefficient was $r = .25$ $p \leq .001$ and $r = -.39$ $p \leq .001$ for positive offsets. We found a similar correlation between the standard deviation of object luminance values: $r = .30$ $p \leq .001$ for negative and $r = -.45$ $p \leq .001$ for positive offsets. No significant correlations between perceptual thresholds and target object areas were observed. However, we note that the highest perceptual thresholds in our results were observed in images with very small objects. In post-test discussions, observers reported selecting specific parts of objects to inform their decisions, these were commonly high-contrast regions near target object boundaries.

4.2. Perceptual Threshold Learning

Since no previous work has addressed the problem of perceptual threshold approximation, we cannot compare our model’s performance to existing solutions. Instead to evaluate the validity of our approach we perform 5-fold cross-validation, reporting average MSE between the predicted and ground truth thresholds for our validation set. We first develop a psychometrics-inspired method for finding our model’s decision boundary, which will serve as a threshold to be compared against empirical thresholds from our experiments. This is done by calculating the soft $F1$ score for each of the two suprathreshold classes between the ground truth mask and model prediction for a range of values of x and placing a threshold at the point when $F1$ score

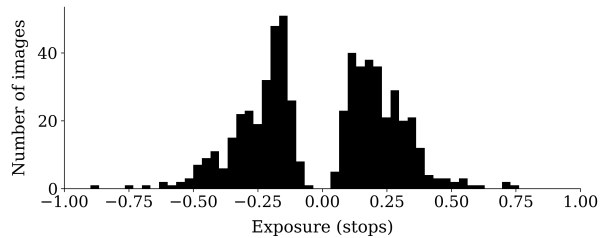


Figure 5. Empirical thresholds collected in our experiment

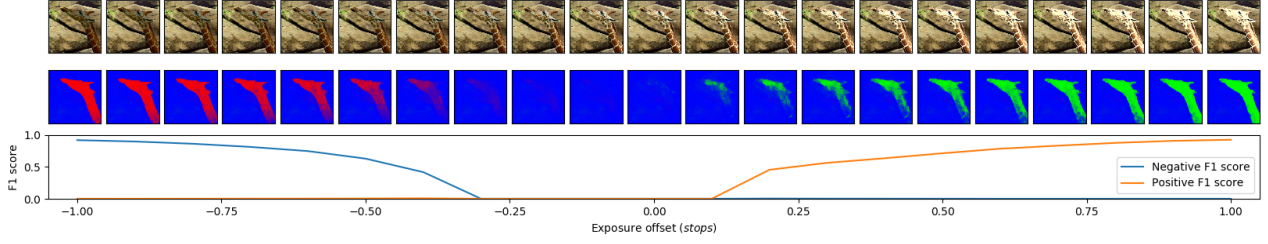


Figure 6. Illustration of how change in $F1$ score between predicted and ground truth (not shown here) masks is used to estimate our model’s decision boundary. The top row shows input images, the middle row shows model prediction softmax probabilities with **red** for detected negative offsets (class 0), **green** for positive offsets (class 1) and **blue** for no offset. The bottom row shows class-wise $F1$ scores for classes 0 and 1. More examples can be found in supplementary materials.

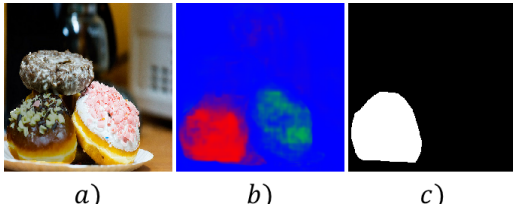


Figure 7. Example of *a)* Over-exposure resulting from flash or spot lighting in the original image *b)* both the original over-exposure (**green**) and manually applied underexposure (**red**) are detected by our model *c)* mask showing area where negative exposure shift is manually applied

becomes nonzero. In our experiments we use $F1 = 0.1$, see Figure 6 for an illustration of the soft $F1$ score as a function of exposure shift. More visual examples can be found in the supplementary materials.

To evaluate the relevance of features learned by the AET, we perform this analysis for a range of fine-tuning regimes, where different parts of the model are frozen before training. The results of this experiment can be seen in Table 1. Overall, our results indicate the benefits of adopting both the AET and multiscale extension, particularly considering the performance increase afforded by freezing the entire encoder and only fine-tuning the decoder. The model’s performance drops significantly when the pre-training stage is omitted or when all layers of the pre-trained model are allowed to be fine-tuned.

5. Conclusions, Limitations and Future Work

We have presented a novel methodology for the detection of local suprathreshold image transformations based on approximating the function performed by an observer. This is achieved by training a fully convolutional image classifier and conditioning its class decision boundaries using a data generation scheme based on empirical perceptual thresholds corresponding to JNDs. We find that the threshold distributions generated by our model approximate the empirical threshold distributions from our experiments. We

Freeze Up To Layer	MSE both	MSE x_{t-}	MSE x_{t+}
no freeze	3.9690	3.5716	4.3664
block1 pool	0.3028	0.2618	0.3442
block2 pool	0.2098	0.2188	0.2000
block3 pool	0.1895	0.1633	0.2161
block4 pool	0.2350	0.2025	0.2681
block5 pool	0.1335	0.1624	0.1046
concatenate	0.1148	0.1307	0.0978

Table 1. Cross-validation results: Average mean squared validation errors between ground truth thresholds and model predictions are given in exposure stops. Individual errors for positive and negative exposure offsets are shown in the rightmost two columns. Errors in each row are a result of freezing progressive parts of the pre-trained AET backbone.

also confirm that adopting the unsupervised AET approach achieves consistently lower errors than training directly on the empirical data without pre-training. Our method can be applied to a range of local distortions or transformations, such as color shifts, blur, aliasing or subsampling, as long as they can be represented by a transformation and mask. Aside from transformations applied manually, our model detects pre-existing over-exposure in our validation set (see Fig. 7). Our results are constrained by the 8-bit dynamic range of images used in our study and the inherent biases associated with individual observers. However, they show that using CNN architectures and an AET unsupervised pre-training strategy is an efficient method of detecting local transformations in images. While a further detailed study and fine-grained optimization are required to maximize performance, our methodology is effective at approximating perceptual thresholds with respect to a local image transformation. We are currently performing an extended study of our approach against different backbone architectures, training regimes, and optimization strategies. We also intend to apply our methodology as the first stage in automatic composite quality improvement.

References

- [1] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.
- [2] Peter GJ Barten. *Contrast sensitivity of the human eye and its effects on image quality*, volume 21. Spie optical engineering press Bellingham, WA, 1999.
- [3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- [4] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. Citeseer, 2017.
- [5] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.
- [6] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015.
- [7] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2017.
- [8] Andrew P Bradley. A wavelet visible difference predictor. *IEEE Transactions on Image Processing*, 8(5):717–730, 1999.
- [9] Patrick Cavanagh. The artist as neuroscientist. *Nature*, 434(7031):301, 2005.
- [10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.
- [11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- [12] Ioan Cleju and Dietmar Saupe. Evaluation of supra-threshold perceptual metrics for 3d models. In *Proceedings of the 3rd symposium on Applied perception in graphics and visualization*, pages 41–44, 2006.
- [13] Scott J Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, volume 1666, pages 2–15. International Society for Optics and Photonics, 1992.
- [14] Alan Dolhasz, Maite Frutos-Pascual, and Ian Williams. [poster] composite realism: Effects of object knowledge and mismatched feature type on observer gaze and subjective quality. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 9–14. IEEE, 2017.
- [15] Alan Dolhasz, Ian Williams, and Maite Frutos-Pascual. Measuring observer response to object-scene disparity in composites. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 13–18. IEEE, 2016.
- [16] Jaroslav Dusek and Karel Roubíř. Testing of new models of the human visual system for image quality evaluation. In *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, volume 2, pages 621–622. IEEE, 2003.
- [17] Gustav Theodor Fechner, Davis H Howes, and Edwin Garriques Boring. *Elements of psychophysics*, volume 1. Holt, Rinehart and Winston New York, 1966.
- [18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.
- [19] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.
- [20] Recommendation ITU-R BT. 500-11, methodology for the subjective assessment of the quality of television pictures”. *International Telecommunication Union, Tech. Rep.*, 2002.
- [21] Jeffrey P Johnson, Elizabeth A Krupinski, Michelle Yan, Hans Roehrig, Anna R Graham, and Ronald S Weinstein. Using a visual discrimination model for the detection of compression artifacts in virtual pathology images. *IEEE transactions on medical imaging*, 30(2):306–314, 2010.
- [22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [23] Elizabeth A Krupinski, Jeffrey Johnson, Hans Roehrig, John Nafziger, Jiahua Fan, and Jeffery Lubin. Use of a human visual system model to predict observer performance with crt vs lcd display of images. *Journal of Digital Imaging*, 17(4):258–263, 2004.
- [24] Yung-Kai Lai and C-C Jay Kuo. A haar wavelet approach to compressed image quality measurement. *Journal of Visual Communication and Image Representation*, 11(1):17–40, 2000.
- [25] Guanbin Li and Yizhou Yu. Visual saliency based on multi-scale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015.
- [26] Guanbin Li and Yizhou Yu. Contrast-oriented deep neural networks for salient object detection. *IEEE transactions on neural networks and learning systems*, 29(12):6038–6051, 2018.
- [27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [29] Weisi Lin and C-C Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of visual communication and image representation*, 22(4):297–312, 2011.

- [30] Hantao Liu and Ingrid Heynderickx. Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(7):971–982, 2011.
- [31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [32] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [33] Anush K Moorthy and Alan C Bovik. Perceptually significant spatial pooling techniques for image quality assessment. In *Human Vision and Electronic Imaging XIV*, volume 7240, page 724012. International Society for Optics and Photonics, 2009.
- [34] Alexandre Ninassi, Patrick Le Callet, and Florent Autrusseau. Pseudo no reference image quality metric using perceptual data hiding. In *Human vision and electronic imaging XI*, volume 6057, page 6057. International Society for Optics and Photonics, 2006.
- [35] Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II–169. IEEE, 2007.
- [36] Yuri Ostrovsky, Patrick Cavanagh, and Pawan Sinha. Perceiving illumination inconsistencies in scenes. *Perception*, 34(11):1301–1314, 2005.
- [37] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51(1):195–203, 2019.
- [38] Eli Peli. Contrast in complex images. *JOSA A*, 7(10):2032–2040, 1990.
- [39] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Color image database tid2013: Peculiarities and preliminary results. In *European workshop on visual information processing (EUVIP)*, pages 106–111. IEEE, 2013.
- [40] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [41] Alan R Robertson. The cie 1976 color-difference formulae. *Color Research & Application*, 2(1):7–11, 1977.
- [42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [43] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.
- [44] Walter J Scheirer, Samuel E Anthony, Ken Nakayama, and David D Cox. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1679–1686, 2014.
- [45] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.
- [46] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128, 2005.
- [47] Hamid R Sheikh, Zhou Wang, Lawrence Cormack, and Alan C Bovik. Live image quality assessment database release 2 (2005), 2005.
- [48] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 657–665, 2015.
- [49] Ran Shi, King Ngi Ngan, Songnan Li, Raveendran Paramesran, and Hongliang Li. Visual quality evaluation of image object segmentation: Subjective assessment and objective measure. *IEEE Transactions on Image Processing*, 24(12):5033–5045, 2015.
- [50] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.
- [51] Christian J Van den Branden Lambrecht and Olivier Verscheure. Perceptual quality measure using a spatiotemporal model of the human visual system. In *Digital Video Compression: Algorithms and Technologies 1996*, volume 2668, pages 450–461. International Society for Optics and Photonics, 1996.
- [52] Cuong T Vu, Eric C Larson, and Damon M Chandler. Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. In *2008 IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 73–76. IEEE, 2008.
- [53] Thomas SA Wallis and Peter J Bex. Image correlates of crowding in natural scenes. *Journal of Vision*, 12(7):6–6, 2012.
- [54] Ching-Yang Wang, Shiuh-Ming Lee, and Long-Wen Chang. Designing jpeg quantization tables based on human visual system. *Signal Processing: Image Communication*, 16(5):501–506, 2001.
- [55] Zhou Wang and Qiang Li. Video quality assessment using a statistical model of human visual speed perception. *JOSA A*, 24(12):B61–B69, 2007.
- [56] Andrew B Watson and Denis G Pelli. Quest: A bayesian adaptive psychometric method. *Perception & psychophysics*, 33(2):113–120, 1983.
- [57] Felix A Wichmann and N Jeremy Hill. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8):1293–1313, 2001.
- [58] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene

recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

- [59] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on Graphics (TOG)*, 31(4):84, 2012.
- [60] Aron Yu and Kristen Grauman. Just noticeable differences in visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2416–2424, 2015.
- [61] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
- [62] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019.
- [63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

Towards Unsupervised Image Harmonisation

Alan Dolhasz^a, Carlo Harvey^b and Ian Williams^c

Digital Media Technology Lab, Birmingham City University, Birmingham, UK
{alan.dolhasz, carlo.harvey, ian.williams}@bcu.ac.uk

Keywords: image compositing, harmonisation, artifact detection, end-to-end compositing, deep learning

Abstract: The field of image synthesis intrinsically relies on the process of image compositing. This process can be automatic or manual, and depends upon artistic intent. Compositing can introduce errors, due to human-detectable differences in the general pixel level transforms of component elements of an image composite. We report on a pilot study evaluating a proof-of-concept automatic image composite harmonisation system consisting of a state-of-the-art deep harmonisation model and a perceptually-based composite luminance artifact detector. We evaluate the performance of both systems on a large data-set of 68128 automatically generated image composites and find that without any task-specific adaptations, the end-to-end system achieves comparable results to the baseline harmoniser fed with ground truth composite masks. We discuss these findings in the context of extending this to an end-to-end, multi-task system.


1 INTRODUCTION


Image compositing is a common task in image processing where an *object* from one image is extracted and inserted into another image, referred to as the *scene*, with the aim of creating a plausible, realistic result (Wright, 2013). Due to inherent disparities in appearance between the object and scene, commonly resulting from differences in illumination, camera intrinsics, post-processing, encoding or compression, component elements of a composite often require post-processing in order to create a compelling and realistic final result. To address these issues, a wide range of automatic compositing techniques have been proposed. These include alpha matting - linear combinations of object and scene pixel values (Porter and Duff, 1984), gradient-domain optimization techniques (Pérez et al., 2003; Agarwala et al., 2004; Levin et al., 2004), visual appearance transfer (Reinhard et al., 2001; Lalonde and Efros, 2007) and multi-scale methods (Burt and Adelson, 1983a,b; Sunkavalli et al., 2010).


More recently, deep learning (DL) based approaches have achieved considerable success in the domain of image compositing. Notably Tsai et al. (2017) adopt the denoising autoencoder (DAE) (Vincent et al., 2008) attempting to learn the compos-

ing function directly from image data, including semantic information derived from ground truth semantic segmentation labels. Chen and Kae (2019) leverage a generative adversarial network (GAN) to learn both colour-based and geometric transformations in order to perform compositing of arbitrary objects into arbitrary scenes. Conditional GANs have also been adopted to address this problem, by learning to model joint distributions of different object classes and their interactions in image space (Azadi et al., 2018), as well as performing colour and gradient blending between composite elements, which have been semantically aligned.

These existing methods are not without limitations. Firstly, they focus on creation of new composites, thus requiring object/scene segmentation masks to be available at input. This limits their use for cases where these are not available, such as improvement of existing image composites. Secondly, they do not explicitly leverage human perceptual characteristics, such as their sensitivity to various image artifacts or magnitude of mismatch between object and scene (Dolhasz et al., 2016). Finally, the masks supplied to such algorithms provide only a binary indication whether a given pixel belongs to the object or the scene. This implies the entire region requires correction and induces a generic transformation, such as colour transfer, uniformly across the region. This can result in the harmonisation algorithm over-compensating and generating a suboptimal out-

^a  <https://orcid.org/0000-0002-6520-8094>

^b  <https://orcid.org/0000-0002-4809-1592>

^c  <https://orcid.org/0000-0002-0651-0963>

put, even compared to the unprocessed input composite.

We argue that *perceptual detection* of composite artifacts should be *explicitly* modelled in deep-learning-based image compositing and harmonisation. Our reasoning behind this is as follows. Firstly, it enables design of end-to-end harmonisation systems without the need for manually supplied object masks, allowing harmonisation of composites for which masks are not available. Secondly, the explicit encoding of the location and perceptual magnitude of errors in the output of the model allows the process to take advantage of the benefits of multi-task learning in terms of generalisation (Ruder, 2017; Ranjan et al., 2017). The potential applications of such automatic compositing systems are wide-ranging, including improvement of legacy content, detection of image manipulations and forgery, perceptually-based metrics and image synthesis.

Consequently, in this study, we design a proof-of-concept end-to-end compositing pipeline consisting of a *detector* network, which outputs masks corresponding to regions requiring harmonisation, and a *harmoniser* network, which corrects the detected regions. We then assess the impact of object masks predicted by the detector on the accuracy of the harmoniser, compared to using ground truth object masks. Our study adopts two existing networks - the Deep Harmonisation algorithm proposed by Tsai et al. (2017) as the *harmoniser* network, and a perceptually-based fully convolutional network proposed by Dolhasz et al. (2019) as the *detector* network.

We show that our prototype end-to-end system, using the the detector network without any task-specific adaptations or re-training, produces results which are comparable to those obtained using ground truth masks. To our knowledge this is the first work investigating the combination of a deep-learning-based detection model with a composite harmonisation one to both detect and fix composites. We are currently developing a complete, end-to-end version of the model, trained specifically for this purpose.

2 RELATED WORK

2.1 Image Compositing & Harmonisation

Automatic image compositing and harmonisation are both active and challenging problems in the domain of image understanding and processing. Image com-

positing concerns the entire process of combining regions from different source images into a plausible whole, while image harmonisation focuses on the problem of matching the various appearance features between the object and scene, such as noise, contrast, texture or blur, while assuming correctly aligned geometric and illumination properties (Sunkavalli et al., 2010).

Similarly to the problem of image in-painting, compositing and harmonisation are both ill-posed problems (Guillemot and Le Meur, 2013). For a given region requiring correction many different arrangements of pixels could be deemed plausible. This is in contrast to problems where the solution is unique. Depending on the content and context of an image composite, some scene properties, and thus required object corrections, may be inferable from the information contained within the image or its metadata, such as the characteristics of the illuminant (Shi et al., 2016), colour palette, contrast range or the camera response function. Other properties, such as an object’s albedo, texture or shape are often unique to the object and cannot be derived directly from contextual information in the scene. While methods for approximating these do exist (Gardner et al., 2017), they are difficult to integrate into end-to-end systems and can be difficult to parameterise. The recent successes in DL have motivated a number of approaches (Tsai et al., 2017; Azadi et al., 2018; Chen and Kae, 2019) which attempt to exploit the huge amount of natural imagery available in public datasets in order to learn the mapping between a corrupted composite image and a corrected composite, or natural image.

2.2 Multi-task Learning

Due to the abundance of natural image data and the ill-posed nature of the compositing problem, DL approaches are well-suited for this task. However, supervised DL methods require large amounts of annotated data in order to learn and generalise well. This requirement grows along with the complexity of a problem and the desired accuracy. Two popular DL paradigms, unsupervised learning and multi-task learning, are often used to address the issues of data labeling and model generalisation.

In recent years many tasks in image understanding have achieved state-of-the-art performance by incorporating multi-task learning Evgeniou and Pontil (2004), for example in predicting depth and normals from a single RGB image (Eigen and Fergus, 2015), detection of face landmarks (Zhang et al., 2014) or simultaneous image quality and distortion estimation Kang et al. (2015). This is afforded by the implicit

regularization that training a single model for multiple related tasks imposes (Caruana, 1997) and the resulting improved generalisation.

State-of-the-art image harmonisation methods focus largely on improving composites in scenarios where the identity of pixels belonging to the object and scene are known a priori. Tsai et al. (2017) use a DAE-based architecture to map corrupted composites to corrected ones, incorporating a two-task paradigm, which attempts to both correct the composite, as well as segmenting the scene. However, they do not explicitly condition the network to learn anything more about the corruption, such as its magnitude, type or location. Instead they provide location information at input time, using a binary mask. Chen and Kae (2019) uses a similar approach - inputting the object mask at training time, however also introducing mask segmentation and refinement within the GAN, in addition to geometric transformations. The segmentation network, as part of the adversarial training process, discriminates towards ground truth binary masks as an output - omitting any perceptual factor in the discrimination task. This achieves improved results compared to the DAE, however at the cost of a more complex architecture and adversarial training.

Due to the many dimensions along which combinations of object and scene may vary, compositing systems should be equipped to assess such differences before attempting to correct them. Kang et al. (2015) shows that a multi-task approach is an efficient way to ensure that distortions are encoded by the model.

3 METHODOLOGY

3.1 Motivation

Whilst multi-task learning has been shown to be efficient in the coupled process of detecting and correcting arbitrary pixel level transformations within images, perceptually-based encoding of artifacts within masks has not yet been shown to be effective in the image harmonisation field. Before approaching the multi-task model, it is necessary to prove empirically that this end-to-end process is viable. Thus we design an end-to-end approach using two existing standalone networks for both detection and harmonisation to test the efficacy of these perceptual masks in the domain.

3.2 Approach

Our overarching goal is the design of an end-to-end automatic compositing pipeline, capable of detection and correction of common compositing artifacts,

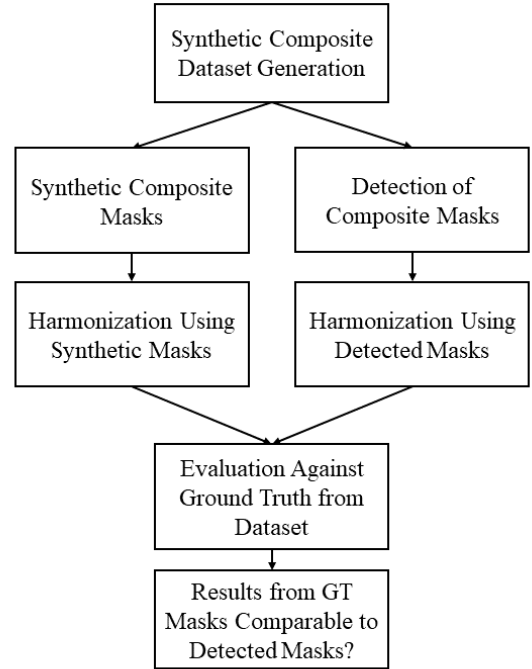


Figure 1: Illustration of research methodology adopted in this work.

without the need for specification of an object mask. In order to evaluate the effectiveness of this approach we propose to assess predicted perceptually-informed object masks rather than ground truth object masks as input to the deep harmonisation algorithm. We then measure similarity between ground truth images and composites corrected with the harmonisation algorithm using either the original synthetic binary masks M_s or the perceptually-based masks predicted by the detector M_p . Accordingly, we refer to composites harmonised using ground truth masks as C_s and composites generated by the end-to-end system as C_p .

We evaluate the hypothesis that the performance of an end-to-end detection and harmonisation model is comparable to a harmonisation model using manually created object masks. Confirmation of this hypothesis would support our case for incorporating explicit detection of composite artefacts into end-to-end image composite harmonisation systems. Our research methodology is summarised in Figure 1.

3.3 Detector and Harmoniser Models

Both the detector (Dolhasz et al., 2019) and the harmoniser (Tsai et al., 2017) are deep, image-to-image, fully convolutional autoencoder networks. The detector takes a single image as input and generates a 2-

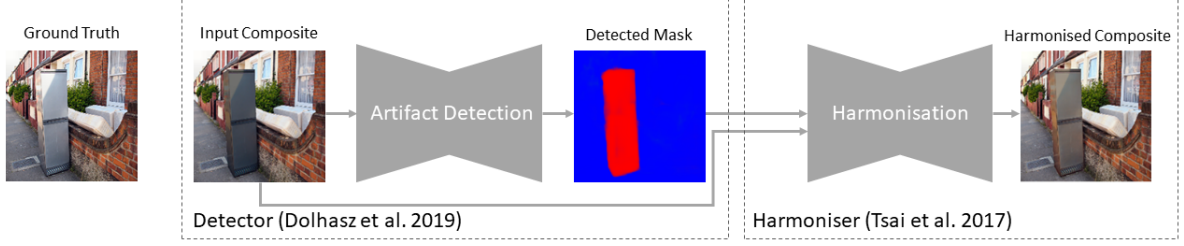


Figure 2: System overview: illustration of the detector and harmoniser combined into an end-to-end composite harmonisation system. A synthetic composite image is first supplied to the detector, which outputs a 2-channel mask indicating detected **negative** and **positive** (not pictured here) luminance shifts. This mask is converted to a single-channel representation by taking a maximum over predicted pixel-wise probabilities and fed to the harmoniser network. The harmoniser then produces a harmonised composite, which we compare against the ground truth.

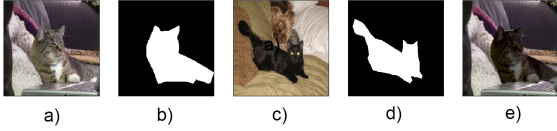


Figure 3: Dataset generation process adapted from Tsai et al. (2017): a) source image sampled from MSCOCO, b) corresponding object mask, c) target image, d) target image object mask, e) result of luminance transfer (Reinhard et al., 2001) of source - c), to target - e

channel output mask, which encodes probabilities for each pixel, p , in the input image as being affected by a negative (channel 0) or a positive (channel 1) perceptually suprathreshold luminance offset. We combine these two suprathreshold channels by taking a pixel-wise maximum $\max(p_0, p_1)$. This way we generate a single mask in the same format as M_s from MSCOCO, where each pixel encodes the probability of a suprathreshold luminance offset. We do not apply any modifications to the harmoniser and adopt the authors’ original trained implementation. The final detector+harmoniser system can be seen in Figure 2.

3.4 Dataset

To perform a fair comparison, we follow the composite generation approach of Tsai et al. (2017). Specifically, we sample pairs of images containing objects belonging to the same semantic category (e.g. person, dog, bottle etc.) from the MSCOCO dataset (Lin et al., 2014). Using their corresponding object masks, we perform statistical colour transfer based on histogram matching, proposed by Reinhard et al. (2001). This process can be seen in Figure 3. This colour transfer is performed between object regions of the same semantic category. As the detector is only conditioned for luminance offsets, we perform colour transfer only

on the luminance channel of Lab colourspace. We generate a total of 68128 composites and corresponding ground truth images. We also extract corresponding the ground truth masks for comparison against the masks predicted by the detector.

3.5 Similarity Metrics

To evaluate each of the two approaches, we calculate similarity metrics between ground truth images C_{gt} and composites corrected by the methods under test: C_s and C_p . We adopt the objective metrics used in the original work, i.e. Mean Squared Error (MSE):

$$MSE = \frac{1}{N} \sum_{i=0}^n (Y_i - \hat{Y}_i)^2 \quad (1)$$

where Y is the ground truth and \hat{Y} is the harmonised image (either C_p or C_s), and Peak Signal-to-Noise ratio (PSNR):

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right) \quad (2)$$

here R is the maximum possible pixel intensity - 255 for an 8 bit image. In addition, we leverage the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), which measures similarity based on human perceptual characteristics. We denote these errors with subscripts referring to the method the composite was fixed with, e.g. MSE_p for MSE between the ground truth image and corresponding composite fixed using predicted masks; MSE_s for MSE between ground truth and a composite fixed using the original MSCOCO masks.

3.6 Procedure

Using our generated composite dataset we first evaluate the harmoniser with ground truth masks. We

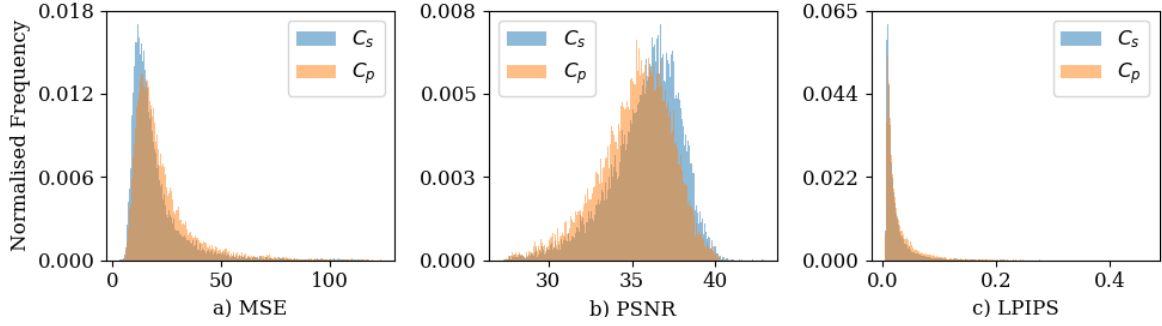


Figure 4: A comparison of the distributions for both C_s (composites corrected with synthetic ground truth masks) and C_p (corrected with masks predicted by the detector) with the number of images in each bin for each metric value. This is shown for: (a) MSE, (b) PSNR and (c) LPIPS. Larger values of MSE and LPIPS indicate poorer performance, whilst this is true for smaller values of PSNR.

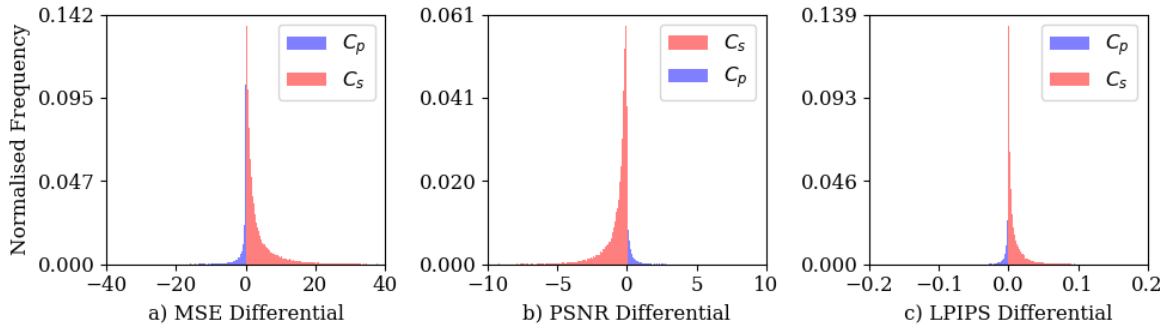


Figure 5: The image-wise error differentials for $C_p - C_s$. This is shown for each of the three metrics: (a) MSE, (b) PSNR and (c) LPIPS. Note, negative values for MSE and LPIPS indicate images for which C_p (composites corrected with masks predicted by the detector) achieves lower error than C_s (composites corrected with synthetic ground truth masks). For PSNR, this is true for positive values.

then use the same dataset to generate predicted object masks using the detector and feed these along with the corresponding composite images to the harmoniser. We obtain two sets of corrected composites: composites corrected using the ground truth masks C_s and composites fixed using masks predicted by the detector C_p . We then calculate similarity metrics between the ground truth images used to generate the composites in the first place, and each of the two sets of corrected images C_s and C_p . These are reported in the following section.

4 RESULTS

The results of our evaluation can be seen in Figure 4, which shows distributions of each of the similarity metrics calculated between ground truth images and composites fixed using C_s and C_p respectively. Mean similarity metrics can be seen in Table 1. Overall, masks predicted by the detector yield higher average errors across all three metrics compared to the ground

truth masks, however the magnitude of these differences is small for each of the metrics. Figure 5 shows distributions of image-wise error differentials for both techniques.

Metric	harmoniser	detector + harmoniser
MSE	19.55	22.65
PSNR	35.81	35.18
LPIPS	0.0227	0.0292

Table 1: Means of similarity metrics for both techniques evaluated against ground truth: harmoniser, and the detector+harmoniser. Lower is better for LPIPS and MSE, higher is better for PSNR.

5 DISCUSSION

Our results indicate that using detected, instead of ground truth object masks can yield comparable results when performing automatic image composite

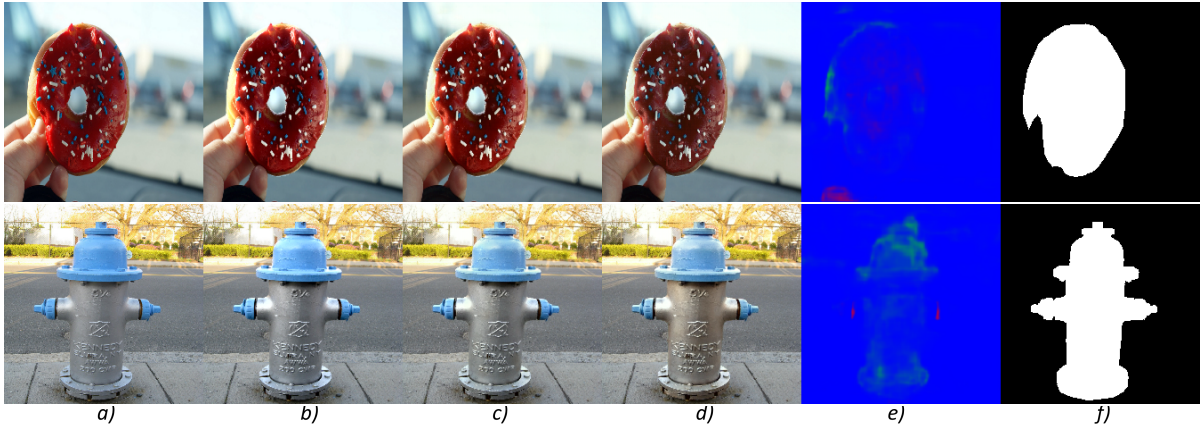


Figure 6: Examples of the harmoniser with ground truth masks over-compensating, and applying colour shifts to compensate a luminance transform, resulting in suboptimal output. From left: *a)* ground truth, *b)* input composite, *c)* output of detector+harmoniser, *d)* output of harmoniser with ground truth masks, *e)* masks predicted by detector, *f)* ground truth masks

harmonisation. Errors obtained using ground truth masks are on average lower compared to those obtained using predicted masks, however in a number of cases the situation is reversed. For example, Figure 6*c* and *d* shows cases of the harmoniser over-compensating, while the detector+harmoniser combination achieves a more natural-looking result. We stress that these results were obtained with no additional training.

Further investigation indicates particular scenarios where this occurs. In some cases, the harmonisation algorithm applies an inappropriate correction, rendering a higher error for C_s compared to the unharmonised input. Then, if M_p does not approximate M_s well, is blank (no detection) or its average intensity is lower than that of M_s , the additional error induced by the harmonisation algorithm is minimised, rendering lower errors for C_p . This can be seen in both images in 6*d*. This indicates the benefit of a perceptually motivated approach to mask prediction, allowing the influence over the weight of the transformation applied by the harmoniser. We also notice that the deep harmonisation network tends to apply colour transformations regardless of whether they are required. In some cases, the perceptually-based masks mitigate this problem. Images showing examples of comparable performance of the two methods can be found in Figure 7. Subfigures *c* and *d* show the results of harmonisation using the approaches under test and subfigures *e* and *f* show M_p and M_s respectively.

Due to the nature of the detector network currently operating solely on luminance transforms, a further benefit to the multi-task learning paradigm is the generalisability to arbitrary pixel level transforms, for example colour shifts. The binary masks accepted by

harmoniser networks currently do not separate across these transforms, they treat them all homogeneously. A perceptually motivated approach to the predicted mask can encode, on a feature-by-feature basis, the perceptual likelihood of harmonisation required. This is not to say necessarily that deep harmonisation networks cannot learn this behaviour, but further support to encode this non-linearity at the input to the network and/or by explicit optimisation at the output, particularly in a multi-task context, would likely benefit performance and improve generalisation (Caruana, 1997).

6 CONCLUSION

These findings, obtained by combination of off-the-shelf models, not modified or re-trained for this specific task, indicate that information about location and magnitude of composite artifacts can be useful in improving the performance of existing compositing and harmonisation approaches. Furthermore, our results show that the requirement for provision of object masks for such algorithms can be relaxed or removed entirely by the explicit combination of composite artifact detection with their correction. This provides a basis for investigation in future work of joint modelling of both the detection and correction of composite image artifacts, e.g. under a multi-task learning paradigm.

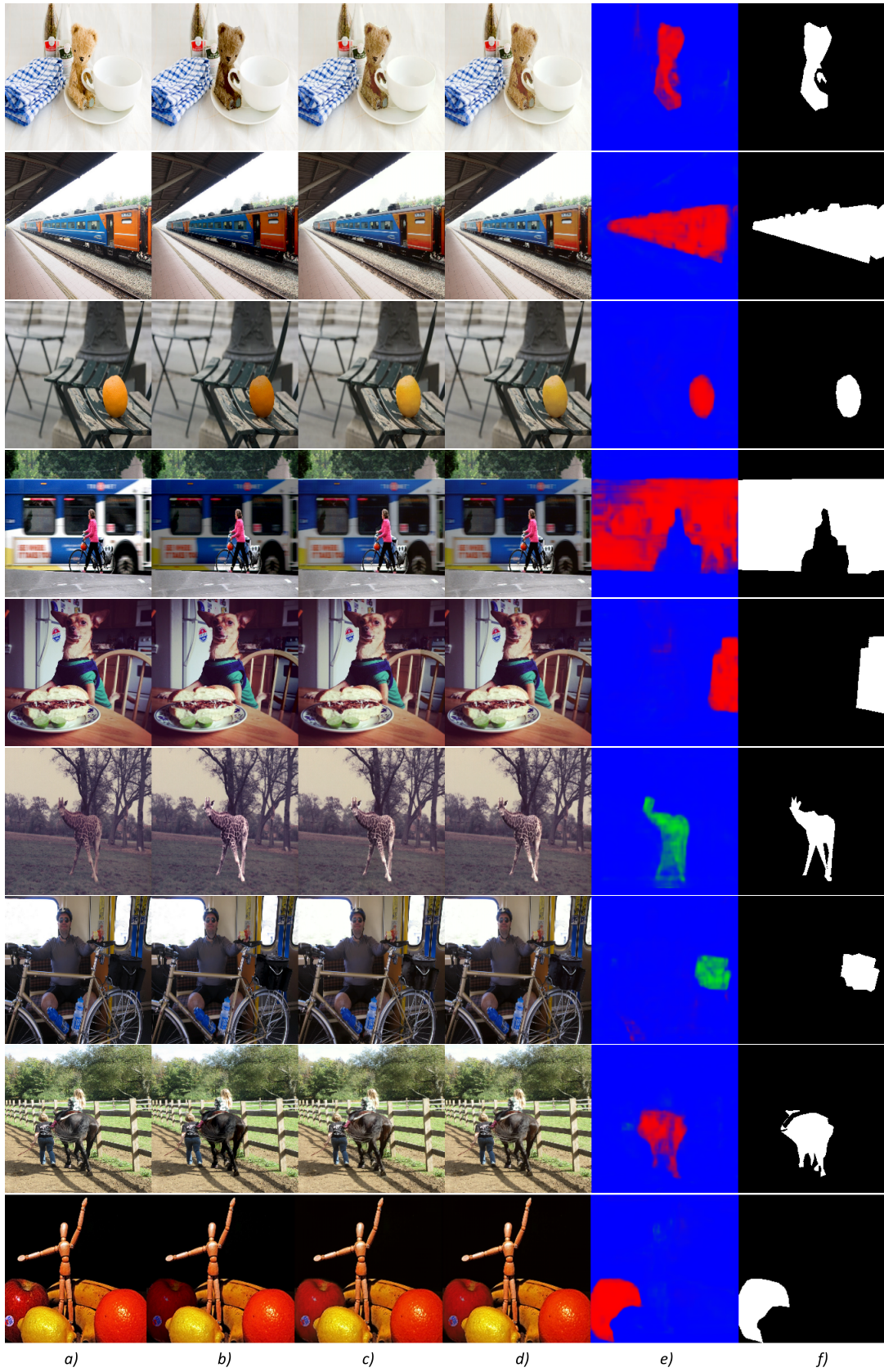


Figure 7: Comparison of harmonisation outputs from our evaluation. From left to right: a) ground truth, b) input composite, c) corrected with detector+harmoniser C_p , d) corrected with ground truth masks + harmoniser C_s , e) Detected masks M_p , f) ground truth masks M_s . Masks in colour indicate the raw output of the detector, where the direction of detected luminance shifts is indicated - red for negative and green for positive shifts.

REFERENCES

- Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., and Cohen, M. (2004). Interactive digital photomontage. In *ACM Transactions on Graphics (ToG)*, volume 23, pages 294–302. ACM.
- Azadi, S., Pathak, D., Ebrahimi, S., and Darrell, T. (2018). Compositional gan: Learning conditional image composition. *arXiv preprint arXiv:1807.07560*.
- Burt, P. and Adelson, E. (1983a). The laplacian pyramid as a compact image code. *IEEE Transactions on communications*, 31(4):532–540.
- Burt, P. J. and Adelson, E. H. (1983b). A multiresolution spline with application to image mosaics. *ACM transactions on Graphics*, 2(4):217–236.
- Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.
- Chen, B.-C. and Kae, A. (2019). Toward realistic image compositing with adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8415–8424.
- Dolhasz, A., Harvey, C., and Williams, I. (2019). Learning to observe: Approximating human perceptual thresholds for detection of suprathreshold image transformations. *arXiv preprint arXiv:1912.06433*.
- Dolhasz, A., Williams, I., and Frutos-Pascual, M. (2016). Measuring observer response to object-scene disparity in composites. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 13–18. IEEE.
- Eigen, D. and Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658.
- Evgeniou, T. and Pontil, M. (2004). Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM.
- Gardner, M.-A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., and Lalonde, J.-F. (2017). Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*.
- Guillemot, C. and Le Meur, O. (2013). Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1):127–144.
- Kang, L., Ye, P., Li, Y., and Doermann, D. (2015). Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *2015 IEEE international conference on image processing (ICIP)*, pages 2791–2795. IEEE.
- Lalonde, J.-F. and Efros, A. A. (2007). Using color compatibility for assessing image realism. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.
- Levin, A., Lischinski, D., and Weiss, Y. (2004). Colorization using optimization. In *ACM transactions on graphics (tog)*, volume 23, pages 689–694. ACM.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Pérez, P., Gangnet, M., and Blake, A. (2003). Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318.
- Porter, T. and Duff, T. (1984). Compositing digital images. In *ACM Siggraph Computer Graphics*, volume 18, pages 253–259. ACM.
- Ranjan, R., Patel, V. M., and Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135.
- Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P. (2001). Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41.
- Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Shi, W., Loy, C. C., and Tang, X. (2016). Deep specialized network for illuminant estimation. In *European Conference on Computer Vision*, pages 371–387. Springer.
- Sunkavalli, K., Johnson, M. K., Matusik, W., and Pfister, H. (2010). Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)*, 29(4):125.
- Tsai, Y.-H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., and Yang, M.-H. (2017). Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797.
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.
- Wright, S. (2013). *Digital compositing for film and video*. Routledge.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.
- Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014). Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer.

Perceptually-Informed No-Reference Image Harmonisation

Alan Dolhasz¹[0000–0002–6520–8094], Carlo Harvey¹[0000–0002–4809–1592], and Ian Williams¹[0000–0002–0651–0963]

Digital Media Technology Lab, Birmingham City University, Birmingham, UK
{alan.dolhasz, carlo.harvey, ian.williams}@bcu.ac.uk
<http://dmtlab.bcu.ac.uk>

Abstract. Many image synthesis tasks, such as image compositing, rely on the process of image harmonisation. The goal of harmonisation is to create a plausible combination of component elements. The subjective quality of this combination is directly related to the existence of human-detectable appearance differences between these component parts, suggesting that consideration for human perceptual tolerances is an important aspect of designing automatic harmonisation algorithms. In this paper, we first investigate the impact of a perceptually-calibrated composite artifact detector on the performance of a state-of-the-art deep harmonisation model. We first evaluate a two-stage model, whereby the performance of both pre-trained models and their naive combination is assessed against a large data-set of 68128 automatically generated image composites. We find that without any task-specific adaptations, the two-stage model achieves comparable results to the baseline harmoniser fed with ground truth composite masks. Based on these findings, we design and train an end-to-end model, and evaluate its performance against a set of baseline models. Overall, our results indicate that explicit modeling and incorporation of image features conditioned on a human perceptual task improves the performance of no-reference harmonisation algorithms. We conclude by discussing the generalisability of our approach in the context of related work.

Keywords: image compositing · harmonisation · artifact detection · end-to-end compositing · deep learning

1 Introduction

Image harmonisation is an important task in image compositing and synthesis, aiming to minimise appearance-based differences between individual elements of a composite, in order to produce a perceptually plausible end result [32]. An image composite commonly consists of at least one *object*, inserted into a background image, referred to as the *scene*. As the object and scene are commonly captured under different environmental conditions, visible appearance mismatches between them may exist, due to differences in illumination, camera intrinsics, post-processing, encoding or compression. Thus, the goal of image

harmonisation is to minimise such differences and create a realistic result. This process can be performed manually by compositing artists, however, many automatic approaches have been proposed, including alpha matting - linear combinations of object and scene pixel values [23], gradient-domain optimization techniques [22, 1, 20], statistical appearance transfer [25, 19] and multi-scale methods [4, 5, 29].

With the advent of deep learning (DL), automatic image synthesis techniques have garnered renewed interest and afforded considerable improvements in state-of-the-art image compositing and harmonisation techniques. Methods using variants of convolutional autoencoders (AEs) have been successfully used to directly approximate the harmonisation function, in a supervised learning setting. Notably, Tsai et al. (2017) [30] use a convolutional AE in a multi-task setting to both segment and harmonise an input image, provided the target object mask. Another approach [7] uses a generative adversarial network (GAN) to perform both colour and geometric transformations, pre-training their model on synthetically-generated data. Conditional GANs have also been applied in this context, by learning to model joint distributions of different object classes and their relationships in image space. This allows for semantically similar regions to undergo similar transformations [2]. A more recent method combines state-of-the-art attention mechanisms and GAN-based architectures with explicit object-scene knowledge implemented through masked and partial convolutions and provide a dedicated benchmark image harmonisation dataset, dubbed iHarmony [8].

A common requirement of these state-of-the-art techniques is the provision of binary object/scene segmentation masks at input, both during training and inference. These masks serve as an additional feature, identifying the corresponding image pixels that require harmonisation. As such, these methods are applicable to scenarios where new composites are generated, and these masks are available. However, in cases where these ground truth masks are not available, these techniques can not be easily applied without human intervention, limiting their application to scenarios such as harmonisation of legacy composites. Moreover, existing methods do not explicitly leverage human perception - the usual target audience of image composites. This includes human sensitivity to different local image disparities between object and scene, shown to correlate with subjective realism ratings [12]. Lastly, binary object masks used in these techniques provide only limited information about the nature of the required corrections, indicating only the area where corrections are needed. This can result in the harmonisation algorithm over- or under-compensating in different local regions of the composite.

In a recent pilot study [11], the authors argue that explicit modeling of the *perception* of compositing artifacts, in addition to their improvement, would allow for harmonisation algorithms to be used in a *no reference* setting, whereby the input mask is not required at inference time. Thus, the model performs both the *detection* and *harmonisation* task. They also show that combining two off-the-shelf, pre-trained models - a detector [10] and a harmoniser [30] - can

achieve comparable results to mask-based state-of-the-art harmonisation algorithms. This enables design of end-to-end harmonisation networks without the need for input object masks, allowing automatic harmonisation of content for which masks are not readily available. The authors also claim that the explicit encoding of the location and perceptual magnitude of errors in the model could allow the process to take advantage of the benefits of multi-task learning, feature sharing and attention mechanism in terms of generalisation [24, 26]. The potential applications of such automatic compositing systems are wide-ranging, including improvement of legacy content, detection of image manipulations and forgery, perceptually-based metrics and image synthesis.

In this paper, we recapitulate and extend this work to an end-to-end model designed, trained and evaluated from scratch. First, we present the original proof-of-concept two-stage compositing pipeline [11]. This consists of a *detector* network, which outputs masks corresponding to regions in an input image requiring harmonisation, and a *harmoniser* network, which corrects the detected regions. We then evaluate the performance of the harmoniser based on using object masks predicted by the detector, versus using ground truth object masks. Based on the evaluation of the two-stage model, we then propose a single end-to-end model, and compare its performance to a set of baselines trained from scratch on the challenging iHarmony dataset, as well as the synthetic COCO-Exp dataset from the original study [11]. We show that our end-to-end model outperforms the baselines on both datasets. This indicates the usefulness of the pre-trained perceptual features to the compositing task using two different end-to-end architectures. To our knowledge, this is the first work investigating an end-to-end combination of a DL-based feature extractor, conditioned on a perceptual task, with an image harmonisation network to perform no reference image harmonisation.

The remainder of the paper is structured as follows: Section 2 introduces related work and discusses state-of-the-art techniques, Section 3 describes the original methodology adopted for the two-stage model evaluation, Section 4 presents the results of this evaluation and Section 5 discusses the findings [11]. In Section 6 we detail the methodology, architecture and optimisation details of the proposed end-to-end models, which are evaluated in Section 7. Finally, in Section 8, we review our findings in the context of the original study and wider application to image harmonisation. We also discuss the strengths and weaknesses of our approach, before concluding and considering future research directions in Section 9.

2 Related Work

2.1 Image Compositing & Harmonisation

Automatic image compositing and harmonisation are both active and challenging problems in the domain of image understanding, synthesis and processing. While, image compositing concerns the entire process of combining regions from different source images into a plausible whole, image harmonisation focuses on the problem of matching the various appearance features between the object and

scene, such as noise, contrast, texture or blur, while assuming correctly aligned geometric and illumination properties [29].

Similarly to the problem of image in-painting, compositing and harmonisation are both ill-posed problems [16]. For a given region requiring correction, many different arrangements of pixels could be deemed plausible. This is in contrast to problems where the solution is unique. Depending on the content and context of an image composite, some scene properties, and thus required object corrections, may be inferred from the information contained within the image or its metadata, such as the characteristics of the illuminant [27], colour palette, contrast range or the camera response function. Other properties, such as an object’s albedo, texture or shape are often unique to the object and cannot be derived directly from contextual information in the scene. While methods for approximation of these properties do exist [15], they are difficult to integrate into end-to-end systems and can be challenging to parametrise. The recent successes in DL have motivated a number of approaches [30, 2, 7, 8] which attempt to exploit the huge amount of natural imagery available in public datasets in order to learn the mapping between a corrupted composite image and a corrected composite, or natural image.

2.2 Multi-task Learning, Feature Sharing & Attention

Due to the abundance of natural image data and the ill-posed nature of the compositing problem, DL approaches are well-suited for this task. However, supervised DL methods require large amounts of annotated data in order to learn and generalise well. This requirement grows along with the complexity of a problem and the desired accuracy. In order to tackle this issue, many architectural considerations have been proposed, many of which focus on learning good feature representations, which generalise well between tasks.

Multi-task learning approaches rely on performing multiple related tasks in order to learn better feature representations. In recent years many tasks in image understanding have achieved state-of-the-art performance by incorporating multi-task learning [14], for example in predicting depth and normals from a single RGB image [13], detection of face landmarks [36] or simultaneous image quality and distortion estimation [17]. This is afforded by the implicit regularisation that training a single model for multiple related tasks imposes [6], and the resulting improved generalisation. Feature sharing approaches combine deep features from related domains or tasks in order to create richer feature representations for a given task. This is similar to the multi-task paradigm, however instead of sharing a common intermediate feature representation, features from one or multiple layers of two or more networks are explicitly combined. The Deep Image Harmonisation (DIH) model [30] adopts both these paradigms, by combining the tasks of image segmentation and harmonisation and sharing deep features of both task branches. Finally, attention mechanisms [9] can also be used to learn the relative importance of latent features for different combinations of task and input sample.

2.3 No more masks

State-of-the-art image harmonisation methods focus largely on improving composites in scenarios where the identity of pixels belonging to the object and scene are known a priori. For example, the DIH approach [30] uses a AE-based architecture to map corrupted composites to corrected ones, incorporating a two-task paradigm, which attempts to both correct the composite, as well as segmenting the scene. However, this approach does not explicitly condition the network to learn anything more about the corruption, such as its magnitude, type or location. Instead object location information is explicitly provided at input, using a binary mask. A similar approach [7] inputs the object mask at training time, while also introducing mask segmentation and refinement within a GAN architecture, in addition to learning of geometric transformations of the object. The segmentation network, as part of the adversarial training process, discriminates towards ground truth binary masks as an output - omitting any perceptual factor in the discrimination task. This achieves improved results compared to the AE, however at the cost of a more complex architecture and adversarial training. Due to the many dimensions along which combinations of object and scene may vary, compositing systems should be equipped to encode such differences before attempting to correct them. Kang et al. (2015) [17] show that a multi-task approach is an efficient way to ensure that distortions are appropriately encoded by the model. Other approaches to this problem include self-supervised pre-training to enforce equivariance of the latent representation to certain input transformations [34], which has been used to train perceptually-aligned local transformation classifiers [10], also used in the proposed model.

3 Two-Stage Model: Methodology

3.1 Motivation

Whilst multi-task learning has been shown to be efficient in the coupled process of detecting and correcting arbitrary pixel level transformations within images, perceptually-based encoding of artifacts within masks has not yet been shown to be effective in the image harmonisation field. Before approaching the multi-task model, it is necessary to prove empirically that this end-to-end process is viable. Thus we first design a two-stage approach using two existing standalone networks for both detection and harmonisation to test the efficacy of these perceptual masks in this domain.

3.2 Approach

Our overarching goal is the design of an end-to-end automatic compositing pipeline, capable of detection and correction of common compositing artifacts, without the need for specification of an object mask. In order to evaluate the effectiveness of this approach, we assess predicted, perceptually-informed object

masks, rather than ground truth object masks, as input to the deep harmonisation algorithm. We then measure similarity between ground truth images and composites corrected with the harmonisation algorithm, using either the original synthetic binary masks M_s or the perceptually-based masks predicted by the detector M_p . Accordingly, we refer to composites harmonised using ground truth masks as C_s , and composites generated by the end-to-end system as C_p .

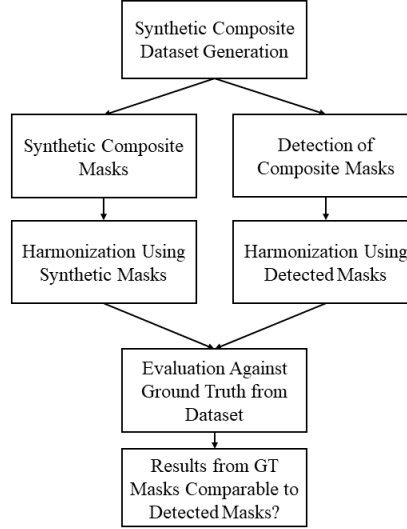


Fig. 1: Illustration of research methodology adopted in the two-stage model evaluation. Reprinted from [11].

We evaluate the hypothesis that the performance of an end-to-end detection and harmonisation model is comparable to a harmonisation model using manually created object masks. Confirmation of this hypothesis would support our case for incorporating explicit detection of composite artefacts into end-to-end image composite harmonisation systems. Our research methodology is summarised in Figure 1.

3.3 Detector and Harmoniser Models

Both the detector (referred to as the PTC henceforth) [10] and the harmoniser (referred to as the DIH) [30] are deep, image-to-image, fully convolutional autoencoder networks. The PTC takes a single image as input and generates a 2-channel output mask, which encodes probabilities for each pixel, p , in the input image as being affected by a negative (channel 0) or a positive (channel 1) perceptually suprathreshold exposure offset. We combine these two suprathreshold channels by taking a pixel-wise maximum $\max(p_0, p_1)$. This way we generate

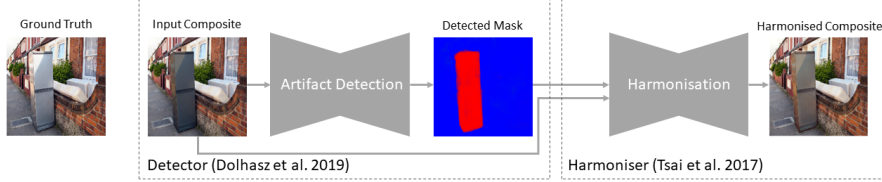


Fig. 2: System overview: illustration of the detector and harmoniser combined into a two-stage composite harmonisation system. A synthetic composite image is first supplied to the detector, which outputs a 2-channel mask indicating detected **negative** and positive (not pictured here) exposure shifts. This mask is converted to a single-channel representation by taking a maximum over predicted pixel-wise probabilities and fed to the harmonisation network, which then produces a harmonised composite, which we compare against the ground truth. Reprinted from [11].

a single mask of the same resolution as M_s , with the difference that each pixel encodes the probability of a suprathreshold exposure offset. We do not apply any modifications to the DIH and adopt the authors’ original trained implementation. The final detector+harmoniser (PTC+DIH) system can be seen in Figure 2.

3.4 COCO-Exp Dataset

To perform a fair comparison, we follow the composite generation approach of [30]. Specifically, we sample pairs of images containing objects belonging to the same semantic category (e.g. person, dog, bottle etc.) from the MSCOCO dataset [21]. Using their corresponding object masks, we perform statistical colour transfer based on histogram matching, proposed by [25]. This process can be seen in Figure 3. This colour transfer is performed between object regions of the same semantic category. As the detector is only conditioned for exposure offsets, we perform colour transfer only on the luminance channel of Lab colourspace. We generate a total of 68128 composites and corresponding ground truth images. We also extract corresponding ground truth masks for comparison against the masks predicted by the detector. For the sake of brevity, we refer to this dataset as *COCO-Exp* throughout the remainder of this paper.

3.5 Similarity Metrics

To evaluate each of the two approaches, we calculate similarity metrics between ground truth images C_{gt} and harmonised images, corrected by the methods under test: C_s (harmonised using ground truth masks), and C_p (harmonised using predicted masks). We adopt the objective metrics used in the original work, i.e. Mean Squared Error (MSE):

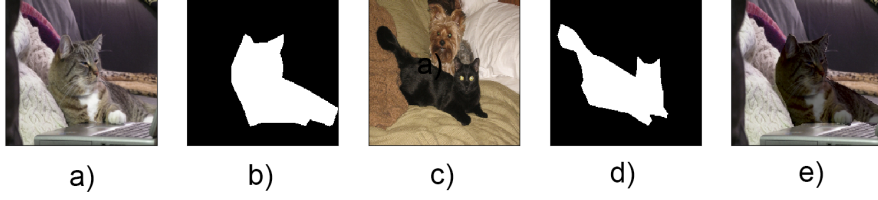


Fig. 3: Dataset generation process adapted from [30]: a) source image sampled from MSCOCO, b) corresponding object mask, c) target image, d) target image object mask, e) result of luminance transfer [25] of source - c), to target - e. Reprinted from [11].

$$MSE = \frac{1}{N} \sum_{i=0}^n (Y_i - \hat{Y}_i)^2 \quad (1)$$

where Y is the ground truth and \hat{Y} is the harmonised image (either C_p or C_s), and Peak Signal-to-Noise ratio (PSNR):

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right) \quad (2)$$

here R is the maximum possible pixel intensity - 255 for an 8 bit image. In addition, we leverage the Learned Perceptual Image Patch Similarity (LPIPS) [35], which measures similarity based on human perceptual characteristics. We denote these errors with subscripts referring to the method the composite was fixed with, e.g. MSE_p for MSE between the ground truth image and corresponding composite fixed using predicted masks; MSE_s for MSE between ground truth and a composite fixed using the original MSCOCO masks.

3.6 Evaluation Procedure

Using our generated composite dataset we first evaluate the DIH with ground truth masks. We then use the same dataset to generate predicted object masks using the PTC and feed these along with the corresponding composite images to the DIH. We obtain two sets of corrected composites: composites corrected using the ground truth masks C_s and composites fixed using masks predicted by the PTC C_p . We then calculate similarity metrics between the ground truth images used to generate the composites in the first place, and each of the two sets of corrected images C_s and C_p . These are reported in the following section.

4 Two-Stage Model: Results

The results of our evaluation can be seen in Figure 4, which shows distributions of each of the similarity metrics calculated between ground truth images and

composites fixed using C_s and C_p respectively. Mean similarity metrics can be seen in Table 1. Overall, masks predicted by the detector yield higher average errors across all three metrics compared to the ground truth masks, however the magnitude of these differences is small for each of the metrics. Figure 5 shows distributions of image-wise error differentials for both techniques.

Metric	DIH	PTC+DIH
MSE	19.55	22.65
PSNR	35.81	35.18
LPIPS	0.0227	0.0292

Table 1: Means of similarity metrics for both techniques evaluated against ground truth: DIH, and the PTC+DIH. Lower is better for LPIPS and MSE, higher is better for PSNR. Reprinted from [11].

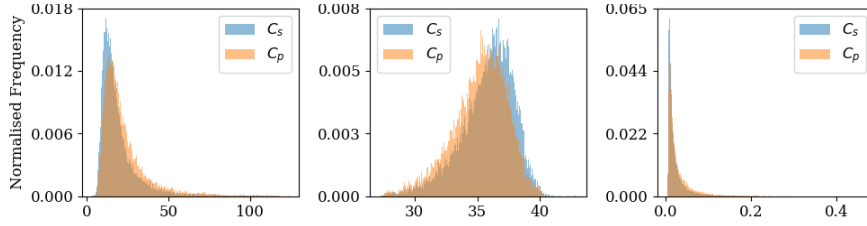


Fig. 4: Similarity metric distributions for both C_s (composites corrected with synthetic ground truth masks) and C_p (corrected with masks predicted by the detector) (a) MSE, (b) PSNR and (c) LPIPS. Larger values indicate poorer performance for MSE and LPIPS, better for PSNR. Reprinted from [11].

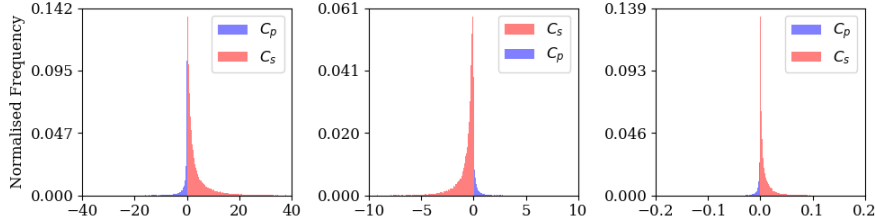


Fig. 5: The image-wise error differentials for $C_p - C_s$, for each of the three metrics: (a) MSE, (b) PSNR and (c) LPIPS. Note, negative values for MSE and LPIPS indicate images for which C_p (composites corrected with masks predicted by the detector) achieves lower error than C_s (composites corrected with synthetic ground truth masks). For PSNR, the obverse is true. Reprinted from [11].

5 Two Stage Model: Discussion

Our results indicate that using detected, instead of ground truth object masks can yield comparable results when performing automatic image composite harmonisation. Errors obtained using ground truth masks are on average lower compared to those obtained using predicted masks, however in a number of cases the situation is reversed. Figure 6 illustrates examples of failure cases, where Figures 6c 6d show cases of the DIH over-compensating, while the PTC+DIH combination achieves a more natural-looking result. We stress that these results were obtained with no additional training. Further investigation indicates particular scenarios where this occurs. In some cases, the harmonisation algorithm applies an inappropriate correction, rendering a higher error for C_s compared to the un-harmonised input. Then, if M_p does not approximate M_s well, is blank (no detection) or its average intensity is lower than that of M_s , the additional error induced by the harmonisation algorithm is minimised, rendering lower errors for C_p . This can be seen in both images in 6d. This indicates the benefit of a perceptually motivated approach to mask prediction, allowing the influence over the weight of the transformation applied by the harmoniser. We also notice that the deep harmonisation network tends to apply colour transformations regardless of whether they are required. In some cases, the perceptually-based masks mitigate this problem. Images showing examples of comparable performance of the two methods can be found in Figure 7. Subfigures *c* and *d* show the results of harmonisation using the approaches under test and subfigures *e* and *f* show M_p and M_s respectively.

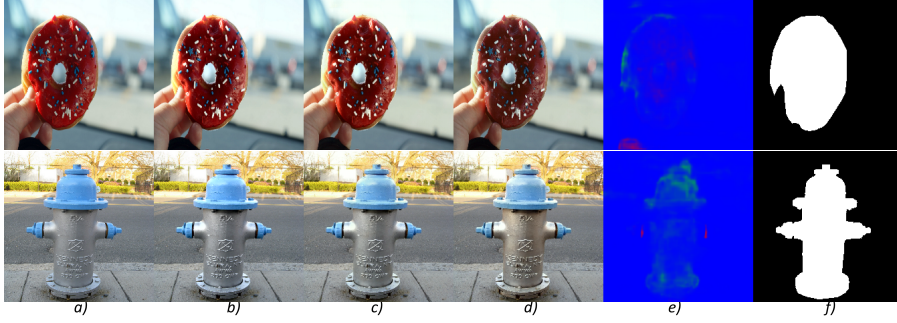


Fig. 6: Examples of the DIH with ground truth masks over-compensating, and applying colour shifts to compensate a luminance transform, resulting in sub-optimal output. From left: *a*) ground truth, *b*) input composite, *c*) output of PTC+DIH, *d*) output of DIH with ground truth masks, *e*) masks predicted by PTC, *f*) ground truth masks. Reprinted from [11].

Due to the nature of the PTC currently operating solely on luminance transforms, a further benefit to the multi-task learning paradigm is the generalisability

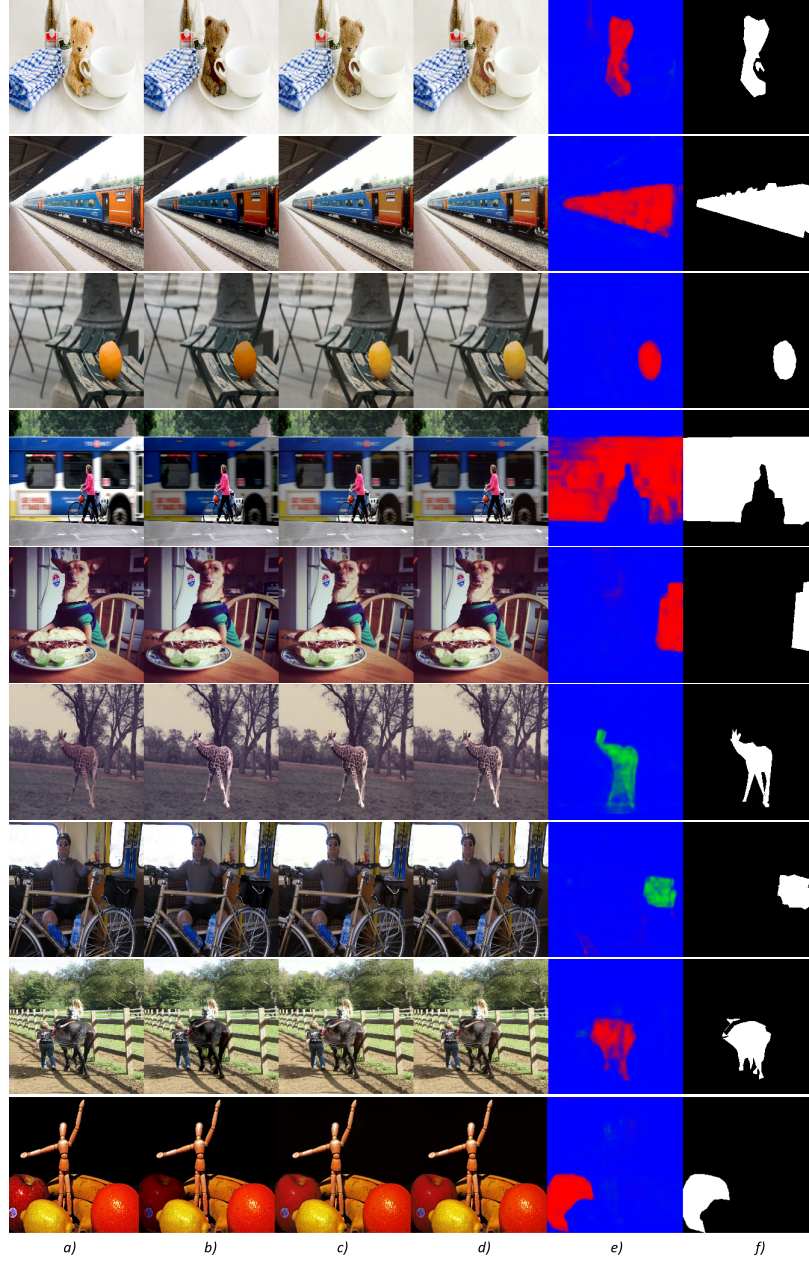


Fig. 7: Comparison of harmonisation outputs from our evaluation. From left to right: a) ground truth, b) input composite, c) corrected with PTC+DIH (C_p), d) corrected with ground truth masks + DIH (C_s), e) Detected masks (M_p), f) ground truth masks (M_g). Masks in colour indicate the raw output of the PTC, where the direction of detected luminance shifts is indicated - **red** for negative and **green** for positive shifts. Reprinted from [11].

to arbitrary pixel level transforms, for example colour shifts. The binary masks accepted by harmoniser networks currently do not separate across these transforms, they treat them all homogeneously. A perceptually motivated approach to the predicted mask can encode, on a feature-by-feature basis, the perceptual likelihood of harmonisation required. This is not to say, necessarily, that deep harmonisation networks cannot learn this behaviour, but provision of further support to encode this non-linearity at the input to the network, and/or by explicit optimisation at the output, would likely benefit performance and improve generalisation [6]. This is conceptually similar to curriculum learning improving convergence in reinforcement learning problems [3], or unsupervised pre-training techniques improving convergence in general.

6 End-to-End Model: Methodology

In Section 4 we illustrated that perceptually-based detection of local image transformations can be leveraged to generate composite masks, achieving comparable results to ground truth masks when evaluated on an image harmonisation task using a state-of-the-art harmonisation model. This indicates that an end-to-end model combining both these tasks could be used to perform *no reference* harmonisation, removing the need for provision of object masks for both training and inference, as opposed to current state-of-the-art approaches. Joint training would also allow for overall performance improvements and enable different combinations of the source models to be evaluated. Thus, to perform a fair evaluation, we implement the end-to-end model and the state-of-the-art baseline from scratch, and train both on the iHarmony dataset [8].

6.1 Model Architectures

The end-to-end model is designed by combining the DIH and PTC models. First, we implement the DIH model in Tensorflow, according to the authors' specification and perform random initialisation. We remove one outer layer of the DIH model, following [8], in order to accommodate for the lower resolution of the PTC and perform all training using a resolution of 256×256 .

We evaluate two approaches to combining the source models. The first approach, *PTC-DIH* combines the models sequentially, whereby the PTC generates a mask from the input image, which is then concatenated with the input and fed to the DIH model, as illustrated in Figure 2. We replace the original 3-class softmax output of the PTC, and replace it with a single-channel sigmoid output, to match the input of the DIH model. We also add up- and downsampling operations in order to adapt the input image to the 224×224 resolution of the PTC, and its output to the 256×256 input of the DIH.

The second approach, *PTC-att-DIH*, inspired by self-attention mechanisms [31], relies on combining the latent features of both models through an attention-like dot product:

$$a_{joint} = fc_3\left(\sigma(fc_1(a_{ptc})) \cdot fc_2(a_{dih})\right) \quad (3)$$

where a_{ptc} is a vector of flattened activations from the bottleneck layer of the PTC, a_{dih} is a vector of activations from the last convolutional layer of the DIH encoder, fc_n are fully-connected layers with 512 neurons each, and σ is a softmax activation.

In both the PTC-DIH and PTC-att-DIH the encoder of the PTC is frozen during training, as in [10], however in the case of PTC-DIH, the decoder of the PTC is allowed to learn, while in the PTC-att-DIH only the encoder is used. The PTC does not receive any additional supervisory signals, such as ground truth object masks, or scene segmentation, only the end-to-end MSE harmonisation loss.

The performance of our joint model is evaluated against two baselines - the vanilla DIH (without semantic segmentation branch), which requires input masks (*DIH-M*), and a no-mask version of the same model (*DIH-NM*), where masks are not provided as input during training. To ensure a fair comparison, we train all models from scratch, using the same dataset and evaluate their performance on the COCO-Exp dataset from Section 3.4 and the iHarmony validation set. We motivate this by the fact that the original PTC implementation is only conditioned on exposure shifts, so a comparison across both datasets can illustrate the performance for simple exposure shifts (COCO-Exp) versus more complex colour transformations (iHarmony). If the perceptually-based features learned by the PTC generalise well across image features, an improvement should be seen over the naive DIH-NM model when evaluated on both these datasets.

6.2 Optimization Details

All of our models are trained for 50 epochs using the entire training set of the iHarmony dataset, consisting of 65742 training images and evaluated using the validation set, consisting of 7404 validation images. The Adam optimizer [18] with default parameters and an initial learning rate of 0.001 is used. We set the batch size to 32 and enforce a 256×256 resolution. We apply pre-processing to all input images scaling the pixel intensity range from $[0, 255]$ to $[-1, 1]$. For each training run, we select the model minimising validation loss for further evaluation.

7 End-to-End Model: Results

This section presents the evaluation of the proposed models on both the validation set of the iHarmony dataset, as well as the COCO-Exp dataset generated for the preliminary study.

Table 2 shows average MSE and PSNR values for both datasets and each of the models. We find that both of our proposed end-to-end models improve performance on both the iHarmony and COCO-Exp datasets, as compared to

the naive baseline, when performing harmonization with no input mask. This suggests the PTC features are relevant to the image harmonisation task. Overall, the PTC-DIH achieves best performance in harmonisation with no input mask, outperforming the PTC-att-DIH and the DIH-NM baseline.

Model	iHarmony		COCO-Exp	
	MSE	PSNR	MSE	PSNR
DIH-M	89	32.56	201	32.18
DIH-NM	153	30.93	276	31.12
PTC-att-DIH	151	31.02	264	31.37
PTC-DIH	124	31.39	214	31.61

Table 2: Test metrics for all evaluated models, across the two datasets used in our experiments. Lower is better for MSE, higher is better for PSNR. Best results using no input mask in bold. Results for the input-mask-based baseline (DIH-M) shown for reference. Higher is better for PSNR, lower is better for MSE.

Figure 8 illustrates the performance of all models under evaluation for several images from the COCO-Exp dataset. Specifically, in each row the input and ground truth are shown in Figures 8a and 8b respectively. Figures 8c, 8e and 8g show the harmonised outputs of the DIH-NM, PTC-att-DIH and PTC-DIH models respectively, while Figures 8d, 8f and 8h are difference image heatmaps between the input and the harmonised output predicted by each model. These heatmaps provide an illustration of the magnitude, direction and location of the applied correction. Upon inspection of similarity metrics, the harmonised outputs and the difference heatmaps, it can be seen that the PTC-DIH model outperforms both the baseline (DIH-NM) and the latent-space-based combination of both models (PTC-att-DIH). This can be seen clearly when comparing the difference images: the PTC-DIH applies corrections more consistently across the region of the target object, compared to the two alternatives. Figure 9 compares the performance of the PTC-DIH to the mask-based DIH-M model for 3 versions of an input image from iHarmony. It can be noticed that the output of both the PTC-DIH and DIH-M closely follow that of the reference. The area corrected by the PTC-DIH aligns with the ground truth mask. Small differences in the output images can be noted, particularly around edges, where the PTC-DIH sometimes contributes to softness and smearing (e.g. Fig.9e, middle row). We found this was often related to artifacts around the edges of objects and near edges of images produced by the PTC. Nonetheless, despite the lack of input mask, the PTC-DIH achieves consistent and comparable results for each of the image variations and, in some cases, avoids the colour shifts induced by the DIH (e.g. compare columns d) and e) with column c) of Figure 9), as discussed in Section 5.

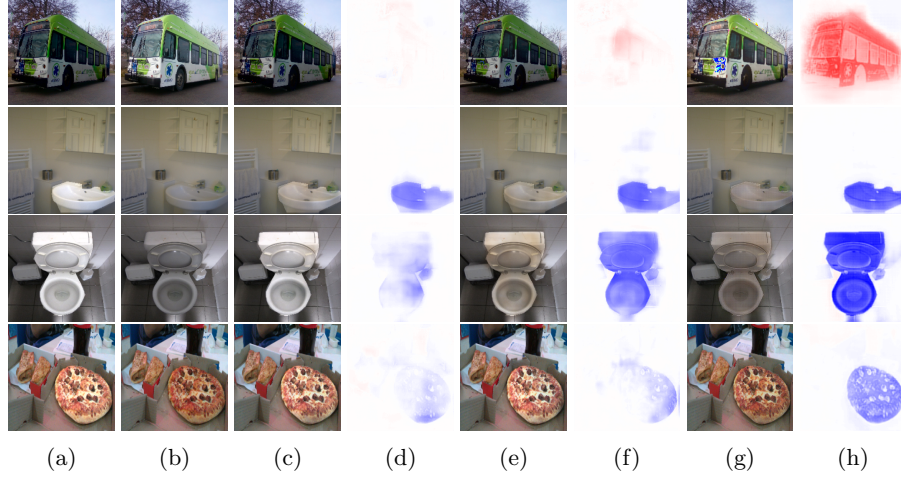


Fig. 8: Comparison of outputs from each model under evaluation for a range of images from the COCO-Exp dataset. *a)* input image *b)* ground truth *c)* DIH-NM result *d)* Difference image between input and output for DIH-NM *e)* PTC-att-DIH result *f)* difference image between input and output for PTC-att-DIH *g)* PTC-DIH result *h)* PTC-DIH difference image. In difference images, red indicates that $\hat{y}_{i,j} - x_{i,j} > 0.0$ whereas blue indicates the opposite.

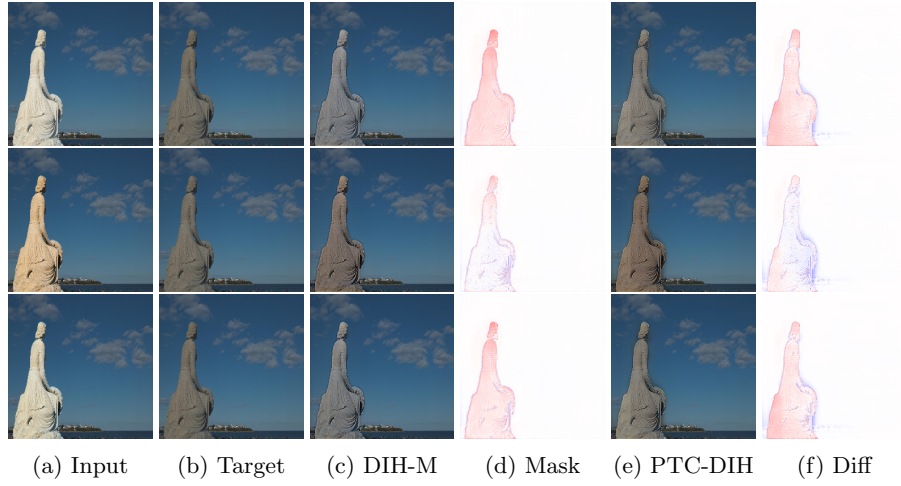


Fig. 9: Comparison between the corrections applied by PTC-DIH, and the mask-based DIH-M models for multiple variants of the same image. *a)* input composite, *b)* ground truth image *c)* output of DIH-M, *d)* Difference heatmap between output of DIH-M and ground truth, *e)* output of PTC-DIH, *f)* Difference heatmap between output of PTC-DIH and ground truth.

Examples of failure cases can be seen in Figure 10. The top two rows illustrate the most common failure case, where the region requiring harmonisation is not detected, and thus not corrected by the model. The top row illustrates this scenario for a larger object size, while the middle row does so for a small object (one of the sheep near the bottom of the image). The bottom row shows a scenario where the harmonisation is performed on the correct object, however the amount of correction is insufficient. In addition, the model applies harmonisation to a part of the image not requiring harmonisation (the screen). This behaviour is likely due to the fact that the PTC was originally conditioned on exposure shifts, resulting in higher sensitivity to over-exposure, compared to other image distortions.

The impact of object size on harmonisation performance of all models is summarised in Table 3 for both the iHarmony and COCO-Exp datasets. Because the MSE is calculated across the entire image, errors are overall lower for smaller objects. However, when comparing the MSE of harmonised images against their baseline MSE (calculated between the input image and ground truth), the relative MSE improvements are greatest for larger objects. This trend is present across both datasets. The PTC-DIH achieves lowest errors in each object size category across both datasets. Notably, for objects in the COCO-Exp dataset with areas ranging 20-40% of the image size, the PTC-DIH model achieves lower errors than the mask-based DIH-M baseline. This illustrates the impact of the PTC being conditioned on only exposure shifts, but also indicates that these features are useful when transferred to a different type of transformations, such as those in iHarmony.

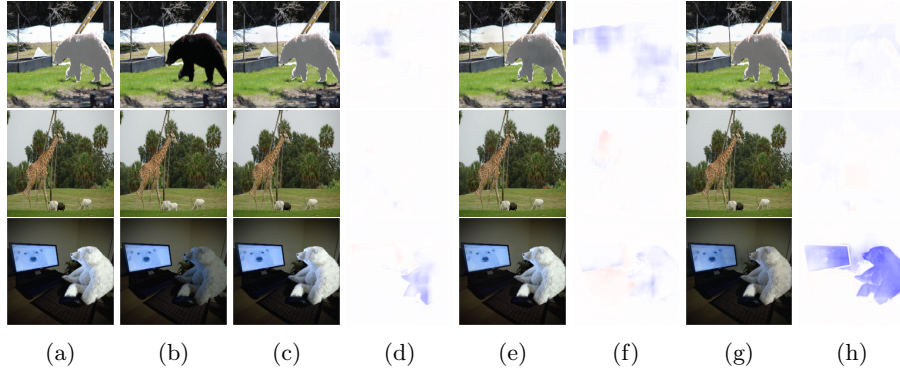


Fig. 10: Examples of failure cases. *a)* input image *b)* ground truth *c)* DIH-NM result *d)* Difference image between input and output for DIH-NM *e)* PTC-att-DIH result *f)* difference image between input and output for PTC-att-DIH *g)* PTC-DIH result *h)* PTC-DIH difference image.

iHarmony								
Object Size	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
DIH-M	33.0	116.1	206.5	335.05	456.2	485.48	484.58	705.12
MSE orig.	47.1	235.02	449.84	642.75	1170.31	1222.97	1151.83	1752.12
DIH-NM	50.73	192.22	360.98	497.42	919.29	1058.39	888.11	1534.94
PTC-att-DIH	50.36	190.2	370.65	462.72	884.22	1001.85	933.02	1659.24
PTC-DIH	45.02	150.04	311.72	359.99	623.03	895.33	720.82	1464.62

COCO-Exp								
Object Size	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
DIH-M	73.74	401.55	655.11	785.35	927.68	1042.68	1119.19	1129.01
MSE orig	86.11	524.29	878.42	1131.53	1503.27	1802.57	2072.08	2097.13
DIH-NM	94.3	502.63	828.69	1045.05	1373.97	1661.55	1876.75	1958.01
PTC-att-DIH	93.26	492.65	802.24	986.49	1271.15	1510.16	1684.99	1806.24
PTC-DIH	82.35	410.08	647.13	778.76	946.99	1084.28	1240.54	1295.38

Table 3: Average MSE on the iHarmony and COCO-Exp datasets for each of the evaluated models, grouped by area of harmonised object as a fraction of image size. *MSE orig* is the MSE between unharmonised inputs and ground truth. Bold values indicate lowest error for each object size, given no mask input. DIH-M model shown for reference.

8 Discussion

The results of both experiments indicate that, in the context of image harmonisation, perceptually-based detection of harmonisation targets can be used to remove the requirement for input object masks. While the proposed approach does not outperform baseline mask-based approaches, it performs significantly better than the state-of-the-art baseline when trained with no input masks. Furthermore, despite the PTC being only conditioned on exposure shifts, its combination with the DIH model improves results on both datasets, suggesting that the perceptually-based features learned by the PTC are useful to the harmonisation task. This is reinforced by the fact that even combining PTC and DIH features in latent space affords a modest improvement over the baseline. Some bias towards exposure shifts is nonetheless noticeable - largest improvements across both datasets occur for achromatic objects (e.g. the sink or toilet in Fig. 8). This could be addressed by training the PTC on a wider range of local transformations. The problem of object size and its impact on harmonisation accuracy is likely connected to the fact that larger objects tend to contribute to the MSE more, compared to smaller objects. The MSE for a small object requiring a 0.5 stop exposure shift will be lower than that of a larger object requiring the same shift. To alleviate this, when training with input masks, the MSE can simply be scaled by the mask size [28], however with no input mask, estimation of target object area becomes nontrivial and presents an interesting direction for further research.

Not unlike the original DIH implementation, the proposed end-to-end model can suffer from gradient artifacts along mask edges, particularly when the initial error to be corrected is large. This issue could be addressed by adopting masked convolutions and utilising self-attention mechanisms, as in [8] or by explicitly incorporating gradient information, as in [33]. While we plan to address these issues in future work, the advantages of our proposed model demonstrated in this work still hold in the context of image harmonisation with no input mask. Following [10], we argue that in order to improve image harmonisation performance, particularly in scenarios where input masks are not available, detection of target regions for harmonisation should leverage intermediate representations equivariant to the transformations of the input to be harmonised. Input masks used in state-of-the-art harmonisation algorithms mimic this role - they encode the presence and location of all input transformations requiring harmonisation as a local binary feature, thus receiving a form of an extra supervisory signal. Our results show that explicitly incorporating the artifact detection paradigm into the harmonisation process can be beneficial, while alleviating the requirements for presence of object masks at inference time.

9 Conclusions & Future Work

In this paper, we have evaluated a novel method for performing image harmonisation without the need for input object masks. Our approach leverages two

state-of-the-art models - an artifact detector and a harmoniser - which, when combined, produce competitive results to mask-based models. We first perform a two-stage evaluation of the original pre-trained models, and based on evaluation results, extend this to a custom end-to-end model in two variants, trained from scratch on the challenging iHarmony dataset. We show that both variants of our end-to-end model outperform the baselines when evaluated on two different datasets. These findings indicate that information about location and magnitude of composite artifacts can be useful in improving the performance of existing compositing and harmonisation approaches. We motivate this by illustrating that ground truth object masks commonly used in harmonisation algorithms essentially substitute the process of detecting local transformations and inconsistencies requiring correction. Accordingly, our results show that the requirement for provision of object masks for such algorithms can be relaxed or removed entirely by the explicit combination of composite artifact detection with their correction. This provides a basis for investigation in future work of joint modeling of both the detection and correction of composite image artifacts, e.g. under a multi-task learning paradigm, where a joint latent representation is conditioned both to be equivariant with respect to input transformations and to encode the structure of the image. In such a scenario, input masks may be used during the training stage, but would not be necessary during inference.

References

1. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M.: Interactive digital photomontage. In: *ACM Transactions on Graphics (ToG)*. vol. 23, pp. 294–302. ACM (2004)
2. Azadi, S., Pathak, D., Ebrahimi, S., Darrell, T.: Compositional gan: Learning conditional image composition. *arXiv preprint arXiv:1807.07560* (2018)
3. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: *Proceedings of the 26th annual international conference on machine learning*. pp. 41–48 (2009)
4. Burt, P., Adelson, E.: The laplacian pyramid as a compact image code. *IEEE Transactions on communications* **31**(4), 532–540 (1983)
5. Burt, P.J., Adelson, E.H.: A multiresolution spline with application to image mosaics. *ACM transactions on Graphics* **2**(4), 217–236 (1983)
6. Caruana, R.: Multitask learning. *Machine learning* **28**(1), 41–75 (1997)
7. Chen, B.C., Kae, A.: Toward realistic image compositing with adversarial learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8415–8424 (2019)
8. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8394–8403 (2020)
9. Cun, X., Pun, C.M.: Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing* **29**, 4759–4771 (2020)
10. Dolhasz, A., Harvey, C., Williams, I.: Learning to observe: Approximating human perceptual thresholds for detection of suprathreshold image transformations.

- In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4797–4807 (2020)
11. Dolhasz, A., Harvey, C., Williams, I.: Towards unsupervised image harmonisation. In: Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP. pp. 574–581. INSTICC, SciTePress (2020). <https://doi.org/10.5220/0009354705740581>
 12. Dolhasz, A., Williams, I., Frutos-Pascual, M.: Measuring observer response to object-scene disparity in composites. In: 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct). pp. 13–18. IEEE (2016)
 13. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
 14. Evgeniou, T., Pontil, M.: Regularized multi-task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 109–117. ACM (2004)
 15. Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.F.: Learning to predict indoor illumination from a single image. arXiv preprint arXiv:1704.00090 (2017)
 16. Guillemot, C., Le Meur, O.: Image inpainting: Overview and recent advances. IEEE signal processing magazine **31**(1), 127–144 (2013)
 17. Kang, L., Ye, P., Li, Y., Doermann, D.: Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: 2015 IEEE international conference on image processing (ICIP). pp. 2791–2795. IEEE (2015)
 18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
 19. Lalonde, J.F., Efros, A.A.: Using color compatibility for assessing image realism. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
 20. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: ACM transactions on graphics (tog). vol. 23, pp. 689–694. ACM (2004)
 21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
 22. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Transactions on graphics (TOG) **22**(3), 313–318 (2003)
 23. Porter, T., Duff, T.: Compositing digital images. In: ACM Siggraph Computer Graphics. vol. 18, pp. 253–259. ACM (1984)
 24. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(1), 121–135 (2017)
 25. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer graphics and applications **21**(5), 34–41 (2001)
 26. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
 27. Shi, W., Loy, C.C., Tang, X.: Deep specialized network for illuminant estimation. In: European Conference on Computer Vision. pp. 371–387. Springer (2016)
 28. Sofiiuk, K., Popenova, P., Konushin, A.: Foreground-aware semantic representations for image harmonization. arXiv preprint arXiv:2006.00809 (2020)

29. Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)* **29**(4), 125 (2010)
30. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3789–3797 (2017)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Advances in neural information processing systems*. pp. 5998–6008 (2017)
32. Wright, S.: *Digital compositing for film and video*. Routledge (2013)
33. Wu, H., Zheng, S., Zhang, J., Huang, K.: Gp-gan: Towards realistic high-resolution image blending. In: *Proceedings of the 27th ACM International Conference on Multimedia*. pp. 2487–2495. ACM (2019)
34. Zhang, L., Qi, G.J., Wang, L., Luo, J.: Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2547–2555 (2019)
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *CVPR* (2018)
36. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: *European conference on computer vision*. pp. 94–108. Springer (2014)